RoboSynChallenge: Mastering Real-World Dexterity via Generalizing Synthesized Manipulation Skills

Runyi Zhao[1,2*†], **Ruixin Wu**[1,2†], **Hongrui Zhang**[1,2†], **Chengkun Li**[1,2†],
Ang Li[2], Ruixing Jin[2], Yueci Deng[2,9], Yingying Guo[2],
Lihe Ding[3], Shaocong Dong[4], Tianfan Xue[3], Yanjun Gao[5], Yudong Luo[6],
Pascal Poupart[7,8], Simo Wu[9], Kui Jia[2,11], Wei-shi Zheng[1,10], Guiliang Liu[1,2]

[1]Shenzhen Loop Area Institute (SLAI) [2] The Chinese University of Hong Kong, Shenzhen
[3] The Chinese University of Hong Kong [4] The Hong Kong University of Science and Technology
[5] LARK Lab, University of Colorado Anschutz [6] Mila - Quebec AI Institute, Canada [7] Vector Institute
[8] University of Waterloo [9] Fudan University [10] Sun Yat-sen University [11] DexForce

robosynchallenge@gmail.com

Abstract

Achieving generalizable robotic manipulation remains a central challenge in embodied intelligence. Despite rapid advances in model architectures and learning algorithms, progress is often limited by the scarcity and narrow diversity of real-world data. The **RoboSynChallenge** competition introduces a unified benchmark to evaluate and advance the *generalizability* of manipulation policies across a spectrum of tasks, environments, and difficulty levels. To alleviate the shortage of realistic data, the challenge integrates large-scale synthetic data generation with standardized real-world robotic evaluation. Participants are encouraged to leverage synthesized state-action trials to improve general-purpose policy learning, while final assessments are conducted exclusively on unseen real-world manipulation environments. Baseline implementations, including Transformer-, Diffusion-, Vision-Language-Action, and World-Action-Model–based policies, are provided to ensure reproducibility and comparability. By coupling scalable simulation-based training with rigorous real-world validation, **RoboSynChallenge** aims to foster the development of broadly capable, data-efficient, and adaptable manipulation systems, thereby paving the way toward truly general robotic intelligence.

Keywords Embodied AI, Sim2Real Transfer, Synthetic Data, Dexterous Manipulation

1 Competition description

Background. Developing scalable and precise robotic manipulation policies represents a critical step toward realizing the vision of *Embodied Artificial Intelligence (EAI)* [1]. As a technology with broad societal and industrial impact, scalable robotic manipulation holds the potential to reduce operational costs and enhance productivity across manufacturing, household, and healthcare domains [2].

Data-driven learning methods have emerged as a principled approach for developing general-purpose robotic agents that can operate flexibly across diverse tasks and environments. In recent years, *generalist policy models* have been advanced through the use of Transformers [3], diffusion models [4], Vision-Language-Action (VLA) frameworks [5], and World Action Models (WAMs) [6, 7, 8, 9]. These architectures enable end-to-end learning pipelines that map multimodal observations and language instructions directly to continuous robot actions [10], paving the way for more autonomous, instruction-following control systems.

*The Leader organizer should be the first author of the proposal.
†Leader and Backup(s) organizers should read and acknowledge Competition Chairs' communications.

Preprint.

As research progresses toward developing generalist manipulation policies, a central question arises: how can we fairly and comprehensively evaluate and benchmark such general-purpose robotic systems? Existing benchmarks are typically categorized as either simulation-based [11, 12, 13, 14] or real-world [15, 16, 17, 18]. However, unlike vision or language models, which can draw upon vast internet-scale datasets, real-world robotic benchmarks rely on data obtained through physical interaction, making data collection costly, slow, and tied to specific hardware setups [5]. While simulation facilitates scalable data generation [19, 20, 21, 22], mismatches in dynamics, kinematics, and sensing inevitably yield a simulation-to-reality (Sim2Real) gap, resulting in degraded real-world performance [14, 23, 24]. Bridging this Sim2Real gap thus remains a fundamental challenge in embodied intelligence research. Moreover, most existing benchmarks rely on fixed datasets, which assume static and controlled environments. This design misaligns with the objective of building generalist policies that must generalize to unseen objects and diverse environments.

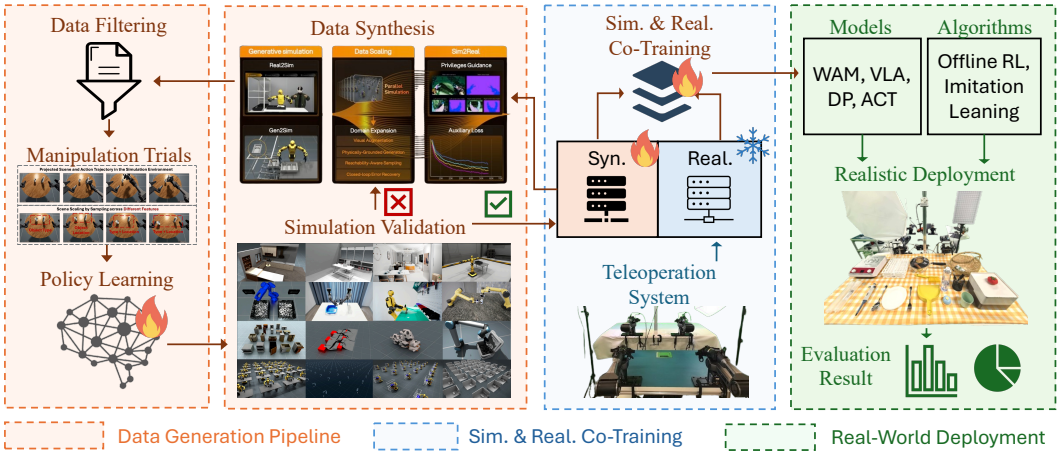

Figure 1: The pipeline of RoboSynChallenge, which provides a comprehensive data generation framework for synthesizing manipulation trials within a simulated environment (in orange background), thereby expanding the training datasets. These synthetic data, together with a smaller amount of static real-world data collected through teleoperation (in blue background), can be used for simulation-to-real co-training. This approach helps mitigate the limitations of scarce real data when deploying trained manipulation policies in the real world (in green background).

Impact. In this proposal, we push the frontier a step further by introducing RoboSynChallenge, a competition designed to quantify *how effectively simulated data can improve the generalizability of manipulation policies when real-world data is scarce*. To achieve this goal, RoboSynChallenge provides a comprehensive data-generation and evaluation pipeline that integrates scalable simulation environments with real-world physical setups for performance validation. The simulation platform enables practitioners to flexibly configure and scale dataset generation across diverse conditions, while the real-world evaluation protocol encompasses multiple levels of difficulty. Together, these components facilitate a rigorous and systematic analysis of Sim2Real transferability.

The RoboSynChallenge is envisioned as an open and scalable benchmark to advance research on generalizable robot learning across simulated and real-world domains. The anticipated impact of RoboSynChallenge is multifaceted as follows:

1) Standardized Benchmark. RoboSynChallenge seeks to establish a standardized real-world benchmark that spans the entire dexterous manipulation policy pipeline, from large-scale data generation and policy learning with synthesized data, to rigorous evaluation under physical environments.

2) Real-world Reproducibility. RoboSynChallenge strengthens the reproducibility of robotic control policies through standardized data synthesis protocols and evaluation frameworks, ensuring consistent benchmarking and fair comparison across real-world environments.

3) Opensource Tools and Baselines. RoboSynChallenge provides publicly available baseline models, data synthesis, and training pipelines, fostering transparent evaluation, reproducible experimentation, and collaborative progress within the robotics research community.

4) Toward Generalizable Manipulation. RoboSynChallenge provides a foundational platform toward generalizable manipulation policies by coupling large-scale synthetic data generation with standardized real-world evaluation, enabling systematic study of scalable policy generalization.

1.1 Novelty

As a new competition, RoboSynChallenge introduces substantial differences compared to previous and ongoing challenges. Our key innovations are as follows:

1) Benchmarking Sim2Real Transferability. RoboSynChallenge aims to fill a critical gap in the field by providing the first standardized benchmark for Sim2Real transferability. It systematically evaluates how well simulated data and policies support scaling to complex, real-world environments. Beyond measuring raw performance, RoboSynChallenge assesses robustness, adaptability, and generalization under real-world noise and domain shifts. This approach fundamentally differs from prior Sim2Sim or Real2Real benchmarks that operate within a single modality. It represents a significant step toward unified Sim2Real evaluation and understanding the limits of simulation-based learning.

2) Generative Data Streaming. Unlike prior benchmarks that rely on a fixed, human-crafted dataset, RoboSynChallenge integrates an automated data generation pipeline [25]. The system procedurally synthesizes simulation environments and automatically generates motion trails in a closed loop. Moreover, to bridge the Sim2Real gap, practitioners can use the provided realistic data as input to style transfer models [26], video generation models [27], or scene augmentation techniques [28, 19] to synthesize diverse motion trajectories with high visual fidelity. This enables large-scale, diverse, and efficient data acquisition without manual curation.

3) Real-World Evaluation. RoboSynChallenge integrates real-world robotic evaluation as part of its official leaderboard. Submitted policies are deployed and tested on standardized physical robot platforms using the same task definitions and evaluation metrics as in simulation. This unified testbed directly measures Sim2Real performance and provides a transparent benchmark for real-world robustness, enabling participants to assess both efficiency and adaptability under tangible constraints.

Table 1 highlights the distinctions between our approach and existing benchmarks. Unlike previous benchmarks that were limited to single-arm settings, simulations, or real-world datasets alone, RoboSynChallenge unifies bimanual manipulation across both simulated and realistic environments. Crucially, it leverages simulated data in a generative manner to augment real-world interactions, enabling richer policy learning and improved generalization. This hybrid paradigm bridges the data efficiency of simulation with the robustness of real-world performance, extending the scope of dexterous manipulation beyond existing competitive and benchmark platforms.

Table 1: The comparison to other competition and benchmarks for dexterous manipulation.

Competition	Rigid Objects	Articulated Objects	Assembling Objects	Tool Using	Manip. Setting	Dataset	Training Env.	Evaluation Env.
RLBench [11]	✓	✓	✗	✗	Single-Arm	Offline	Simulated	Simulated
CALVIN [12]	✓	✓	✗	✗	Single-Arm	Offline	Simulated	Simulated
LIBERO [13]	✓	✓	✗	✗	Single-Arm	Offline	Simulated	Simulated
RoboTwin [29]	✓	✓	✗	✓	Both	Offline	Simulated	Simulated
WBCD [3]	✓	✓	✓	✗	Bimanual	None	Realistic	Realistic
RoboChallenge [15]	✓	✓	✓	✓	Both	Offline	Realistic	Realistic
RobotArena ∞ [16]	✓	✓	✗	✗	Single-Ar	None	Realistic	Realistic
RoboArena [17]	✓	✓	✗	✓	Single-Arm	Offline	Realistic	Simulated
ManipulationNet [18]	✓	✗	✓	✗	Single-Arm	Offline	Realistic	Realistic
ManipArena [4]	✓	✓	✗	✗	Bimanual	Offline	Realistic	Realistic
RoboSynChallenge	✓	✓	✓	✓	Bimanual	Generative	Sim. & Real.	Realistic

1.2 Data

The dataset employed in RoboSynChallenge is primarily generated specifically for this competition using the *EmbodiChain* [25] generative simulation framework. It comprises large-scale embodied interaction trajectories, multi-modal sensory streams (including RGB-D data, proprioceptive signals, and physics-based scene metadata), and structured annotations detailing task semantics, contact events, and success outcomes. To moderate the difficulty of Sim2Real transfer, RoboSynChallenge also includes a smaller subset of data collected through teleoperation to provide real-world correspondence for a limited set of scenarios. The detailed generation process is described as follows.

All robot data were produced within robot manipulation environments using a simulation environment, generative methods, and teleoperation. No personally identifiable or human data are included.

Data Collection Protocol. To facilitate the generalizability of manipulation policy in the RoboSyn-Challenge, we designed protocols for collecting both simulated and real-world datasets.

1) *Real-World Data Collection Protocol.* To construct the real-world dataset, each manipulation task was conducted under five distinct experimental conditions, each further evaluated across four positional variations and three orientation settings, resulting in a total of 60 samples per task. The experimental conditions are designed to introduce variability in background texture, illumination, and scene complexity, thereby enabling a rigorous assessment of generalization in real-world manipulation scenarios. The detailed configuration of the experimental conditions is summarized in Table 2 in Appendix A.1. Each condition was combined with four positional variations and three orientation settings, ensuring comprehensive coverage of spatial and visual factors across all collected samples.

2) *Synthesized Data Collection Protocol.* RoboSynChallengeintegrates Embodichain [25] to scale up the simulated state-action trails. Appendix A.2 shows the details of data generation. Each simulation instance involved systematic perturbations in lighting, object attributes, table properties, background color, and robot configuration. In addition, both intrinsic and extrinsic camera parameters were randomized to simulate different viewpoints and imaging conditions. The randomization schema is summarized in Table 3 in the Appendix. This procedure introduces controlled, multi-level variability, including lighting, materials, geometry, and sensing, which ensures the simulated dataset captures diverse visual and physical conditions for robust policy training and evaluation. We randomly sampled variations to generate 1,000 manipulation trials for each task. The complete data collection pipeline has been released as open-source, *allowing practitioners to flexibly modify and extend the framework according to their specific research needs.*

In addition, since we provide realistic data, practitioners can leverage it to guide video generation models [27] through in-context learning or fine-tuning. The generated data can be further augmented or refined using scene augmentation techniques (e.g., [28, 19]) applied in either simulated or real-world settings. RoboSynChallenge offers flexible integration with such pipelines, facilitating large-scale synthesized data collection.

Confidential Ground Truth and Leakage Prevention. Our evaluation protocol measures the model's real-world performance using diverse metrics (Section 1.4). This evaluation does not require access to the testing data or any ground-truth labels for verification. Instead, it assesses outcomes directly in the real world. Moreover, the training data consist only of successful trajectories collected in the learning environment. To quantify generalizability, the evaluation employs a held-out testing protocol in which the test environments are out-of-distribution relative to the training observations, ensuring minimal overlap and reducing the risk of dataset leakage.

All training and evaluation datasets will be made freely available to registered competition participants through online platforms (e.g., Hugging Face). The dataset is released under a **Creative Commons Attribution–NonCommercial–ShareAlike 4.0 (CC BY-NC-SA 4.0)** license, permitting research use while preventing commercial redistribution. Each release will include *metadata, documentation, and generation code* (for simulated data), in accordance with the NeurIPS Code of Ethics.

1.3 Tasks and application scenarios

Hardware. As Figure 2 shows, the evaluation of manipulation policy relies on a dual-arm platform built from two AgileX Piper manipulators. Each Piper is a 6-DoF research arm featuring open interfaces, designed for embodied manipulation and rapid system integration. This platform has been widely adopted in both prior competitions (e.g., RoboTwin [14], WBCD 2025 [5], and WBCD 2026 [6]) and recent research works [7, 30, 22, 31].

Manipulation tasks. Under the above hardware system, our design tasks are diverse and involve handling rigid objects, articulated objects, deformable objects, and tools. All tasks are performed based solely on visual observations and the robot's proprioceptive feedback. An overview of the real-world setup is shown in Figure 2. The workspace consists of an adjustable-height table with replaceable tablecloth textures and configurable lighting. We categorize the tasks as follows:

1) Entry-level tasks primarily include short-horizon, low-contact-complexity routines with clear affordances. As shown in Figure 3 (green background) and Table 4, they cover table rearrangement, click-bell, water pouring, and handle basket. These tasks emphasize stable perception, reliable grasp execution, and consistent motion trajectories under minimal environmental uncertainty. The focus

[5] https://wbcdcompetition.github.io/2025/index.html

[6] https://wbcdcompetition.github.io

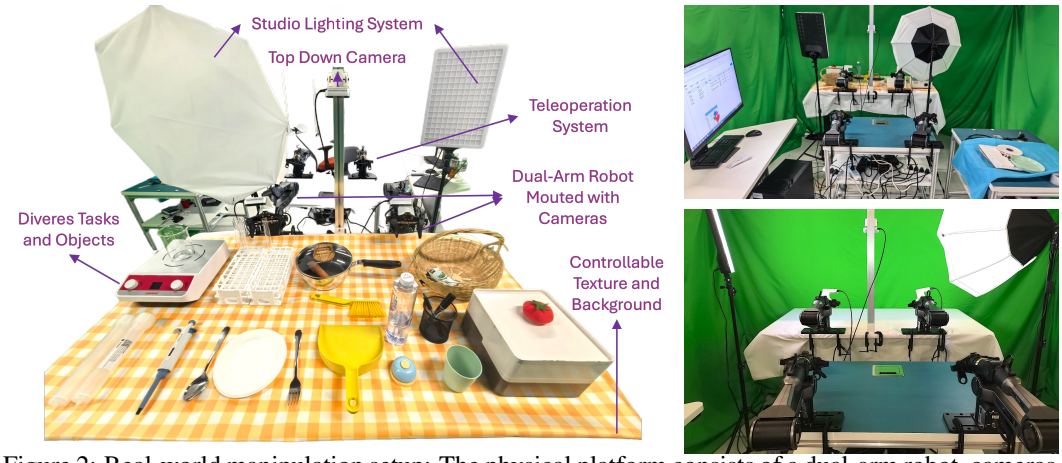

Figure 2: Real-world manipulation setup: The physical platform consists of a dual-arm robot, cameras, lighting, and various objects, enabling evaluation of all tasks included in the benchmark. To facilitate multiple rounds of real-world testing in a reliable and efficient manner, we have prepared *three sets of backup workstations*, each configured identically to the primary setup (Figure 4).

lies in ensuring repeatable performance of simple action primitives rather than complex planning or intricate contact manipulation.

2) Mid-level tasks consist of compound, sequential interactions requiring moderate coordination, adaptive control, and spatial reasoning. As described in Figure 3 (blue background) and Table 4, these include items hand-over, drawer open-and-place, and mixer operating. In contrast to entry-level routines, mid-level tasks introduce variable contact conditions, multi-object dependencies, and partial planning horizons. Evaluation focuses on assessing smooth transition across motion phases, consistent force regulation, and robustness to mild external perturbations.

3) High-level tasks comprise fine-grained manipulation and domain-specific operations demanding high precision, multi-stage sequencing, and dynamic adaptation. As outlined in Figure 3 (red background) and Table 4, they include item assembly, pipette manipulation, and sample loading. These tasks require complex reasoning over ordered procedural steps, precise end-effector control, and environmental feedback integration. Benchmarking evaluates success through accuracy of placement or alignment, procedural consistency, and recovery capability under minor execution deviations.

Across these three levels, RoboSynChallengeprovides in-distribution testing (with respect to real-world training data) for model development and out-of-distribution testing, acting as a held-out benchmark to evaluate generalization quality throughout the competition.

1.4 Metrics

In our experiments, we present the evaluation protocol and metrics as follows.

Evaluation Protocol. We evaluate dexterous manipulation performance across multiple tasks under the following real-world variations: 1) *Background variation*: three table textures (wood, blue fabric, yellow grid). 2) *Lighting variation*: three distinct light positions with varying illumination colors. 3) *Object variation*: both *seen* and *unseen* object instances within the same task category, evaluating generalization to novel appearances. 4) *Distractor presence*: task-irrelevant objects placed in the scene at three difficulty levels (2, 4, and 8 distractors), sampled from a separate object set distinct from task objects. 5) *Spatial generalization*: evaluation on unseen positions within a predefined 3×3 grid. For each task, we first evaluate the policy under a canonical base configuration. We then systematically vary one factor at a time while keeping all other factors fixed. For each configuration, we evaluate the task under multiple object positions by applying small random shifts, and report the resulting success rate across these variations.

Evaluation Metrics. Based on the aforementioned evaluation protocol, we evaluate the performance of a dexterous policy according to the following methods: *1) Success Rate (SR)* (in percentage) measures how often a dexterous manipulation policy completes a task according to predefined success criteria (see Table 4). 2) *Inference Time* measures how quickly the policy produces actionable commands during inference; lighter models typically run faster. All models are deployed on the same machine with an NVIDIA H100 GPU to ensure fair comparison. 3) *Action Steps* count the number

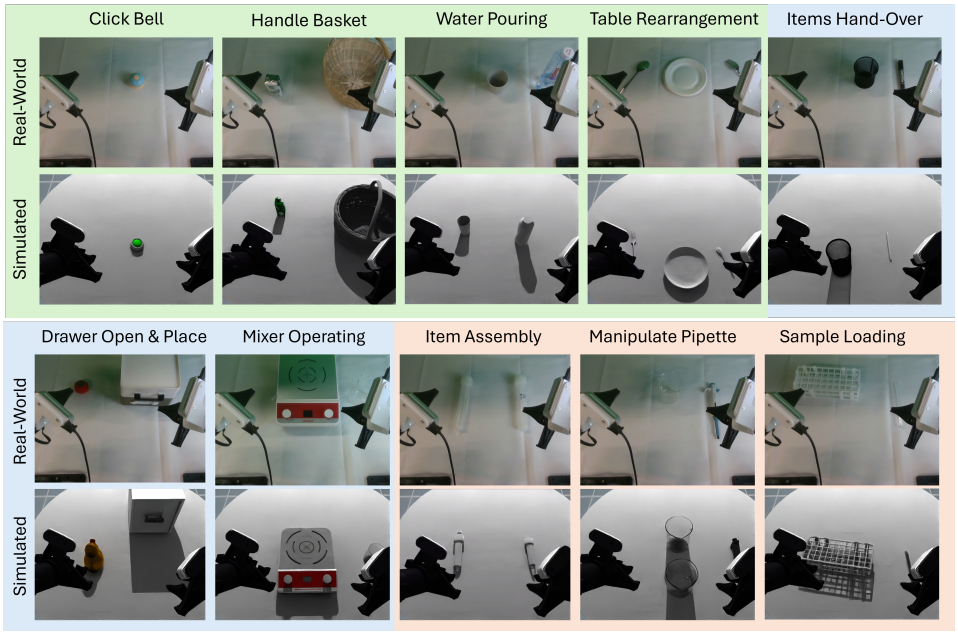

Figure 3: Visualizations of the realistic environments (upper row) and their corresponding simulated environments (lower row) across the three task categories: entry-level (green background), mid-level (blue background), and high-level (red background).

of control steps the model takes to finish a task, such that a higher count can indicate added latency due to error detection, feedback, and resolution, or reflect unresolved failures where the task remains incomplete, and the action budget is exhausted.

1.5 Baselines, code, and material provided

Baselines. The baselines for our task include: 1) $\boldsymbol{\pi_0}$ [32]: A VLA model that combines a flow-matching architecture with a pre-trained vision-language model (VLM). The flow-matching component enables smooth trajectory learning between observations and actions, while the VLM provides strong visual-semantic grounding. 2) $\boldsymbol{\pi_{0.5}}$ [33]: An VLA model that builds upon π_0 by incorporating partial fine-tuning of the pre-trained vision-language backbone within the flow-matching framework. This allows $\pi_{0.5}$ to adapt more effectively to downstream tasks while retaining strong generalization from pre-training. 3) **Motus [7]**: A unified latent action world model that integrates vision, language, and action understanding through a mixture-of-transformers architecture. By leveraging optical-flow-based latent actions and large-scale multimodal data, Motus achieves strong generalization and state-of-the-art performance across both simulated and real-world manipulation tasks. The experimental results for the aforementioned baselines, trained using simulation-only or real-only data, are summarized in the Table 5 in Appendix B below. Furthermore, we plan to incorporate a broader range of state-of-the-art baselines in future work, including VLA models such as **RDT-1B** [34] and WAMs like **DreamZero** [9].

Protocol of Releasing Code and Data We plan to release the starting kit through a dedicated GitHub repository that will contain all necessary materials for participants to easily join the competition. The repository will include: 1) Baseline code for model training and evaluation, 2) Data-loading and preprocessing tools for both simulated and real-world datasets, and 3) Sample simulated data together with a data-generation pipeline to enable participants to reproduce and extend training environments. In addition, we will release the real-world dataset and the complete simulated dataset (with data generator) on a public hosting platform (Hugging Face) to ensure accessibility and scalability.

1.6 Website, tutorial, and documentation

A dedicated competition website[7] will serve as the central hub for all information related to the event. The website will comprehensively present the competition overview, detailed timeline, participation steps, and relevant submission instructions. It will feature a FAQ/Tutorial section offering step-by-step guidance on installing the simulation environment, generating datasets, and testing models in both

[7]Competition website: `https://edem-ai.github.io/robosynchallenge.github.io/`

simulated and real-world settings. To ensure open communication, participants will be able to contact the organizers directly via robosynchallenge@gmail.com, with responses and updates maintained regularly. All content, including documentation and tutorials, will be made available well in advance of the competition start date, and the website will go live within two weeks of acceptance notification. The official GitHub repository [8] will host the codebase, baseline models, and data tools, while an accompanying white paper will describe the competition design, problem formulation, and technical foundation supporting the challenge.

2 Organizational aspects

2.1 Protocol

Steps to Join the Competition. Participants begin by registering on the official competition website, which links to a dedicated GitHub repository containing the starting kit with baseline code, data-loading utilities, and a simulation environment. Teams can set up the environment locally or on cloud platforms using the provided Docker image or setup scripts, then access or generate both real and simulated datasets. They may develop and train models by extending or modifying provided baselines, validate performance in simulation, and ultimately deploy their models through a server-hosted HTTP inference endpoint connected to the competition system. During evaluation, sensory data from real robots are streamed to the participant's endpoint, which returns action commands for real-world execution and benchmarking.

Competition Phases and Platform. The competition unfolds in three sequential phases. Phase I – Simulation Track allows participants to train and submit models within the simulation environment, generating public leaderboard results. Phase II – Real-World Validation invites the top-performing teams to submit inference endpoints for testing on standardized physical robotic systems under controlled conditions. Phase III – Final Evaluation and System Presentation challenges selected teams with unseen tasks, either on-site or remotely, to assess true generalization capabilities, with final results revealed during the closing session.

Cheating and Overfitting Prevention. To ensure fairness and scientific rigor, hidden test sets with varied backgrounds, lighting, and object configurations prevent data leakage and encourage robust generalization. All real-world evaluations are conducted server-side, under identical hardware and environmental conditions, and participants interact only through inference APIs without direct robot access. A single unified model is required across tasks, and intellectual property is protected since only model outputs (not internal code or weights) are transmitted. Submission frequency is limited to curb leaderboard overfitting. Collectively, these mechanisms ensure reproducible and trustworthy results across both simulated and real-world benchmarks.

2.2 Rules and Engagement

Contest Rules. The competition is open to all members of the NeurIPS community and the wider research public, welcoming individuals or teams from academia and industry worldwide. It consists of three phases: a Simulation Track for model development and public evaluation, a Real-World Evaluation for top teams to test models on standardized robotic setups, and a Final Evaluation involving hidden test environments with unseen tasks and real-robot trials.

Participants submit model checkpoints or host inference endpoints according to official API specifications, with limited daily submissions to prevent overfitting. Performance is assessed automatically using standardized metrics such as Success Rate, Inference Time, and Action Steps under consistent hardware and environmental conditions. Hidden test sets with novel variations ensure fairness and true generalization. All participants must adhere to the ethical guidelines and code of conduct.

Communication with Participants. Participants will engage with the organizers and community through openly accessible channels, including a GitHub repository for code releases and technical support, a public discussion forum (in the same GitHub link) for community interaction, and a designated email address for administrative inquiries (see Section 1.6). Official announcements, rule clarifications, and schedule updates will be shared via the competition website, GitHub repository, email notifications, and pinned posts in the public forum. This integrated communication framework ensures transparent, inclusive, and equitable access to information and resources for all participants.

2.3 Schedule and readiness

Timeline competition. Our detailed timeline is presented as follows:

[8] GitHub repository: `https://github.com/EDEM-AI/RoboSynChallenge`

Preparation (May 15 – June 14, 2026): 1) Finalize simulation environments and generate full synthetic datasets. 2) Internal beta testing of the evaluation API and leaderboard.

Launch (June 15, 2026): 1) Official competition launch and website release. 2) Release of "Starting Kit," baseline code (π_0, RDT-1B), and training data.

Development Phase (June 15 – Sept 15, 2026): 1) 90-day window for participants to develop and refine models. 2) Release of "Real-World Sample Set" for Sim2Real alignment. 3) Final submission deadline for model code and endpoints.

Evaluation & Analysis (Sept 16 – Oct 20, 2026): 1) Real-world validation on physical robots for top-performing teams. 2) Verification of reproducibility and final scoring.

Conclusion (Oct 21 – Oct 31, 2026): 1) Publication of final leaderboard and competition summary report. 2) Announcement of winners for NeurIPS 2026 presentation.

2.4 Competition promotion and incentives

RoboSynChallenge executes a multi-channel promotion strategy and a dedicated diversity initiative.

Promotion Plan. We will disseminate the Call for Participation through major academic mailing lists, including robotics-worldwide, ML-news, and the CVPR/ICCV lists. To foster engagement, we will promote the competition via posters and technical webinars showcasing our Starting Kit, as well as by presenting the challenge at leading 2026 robotics conferences such as ICRA and RSS. Our social media outreach will feature "Leaderboard Spotlights" on X (Twitter), LinkedIn, and WeChat, complemented by a featured Hugging Face blog post to engage the broader open-source research.

Incentives and Authorship. Participants will compete for a 20,000 USD prize pool and 5,000 USD in travel grants to attend NeurIPS 2026. The top three teams will receive a podium presentation slot at the NeurIPS workshop. Furthermore, we will lead a joint "findings paper" for a top-tier journal (e.g., JMLR or IJRR). Named authorship will be granted to the top five teams, while all other qualifying participants who outperform the baseline will be credited under a consortium author .

Diversity and Inclusion. To lower participation barriers, all evaluations will be conducted on our side, and a "Zero-Hardware" track will be provided, allowing researchers without access to physical robots to compete via simulation. We will offer cloud computing credits to teams from resource-constrained institutions and conduct targeted outreach to organizations such as Black in AI, LatinX in AI, and WiML. Importantly, 50% of the travel award budget will be reserved for participants from underrepresented groups and researchers from developing nations, ensuring diverse and inclusive representation at the conference.

3 Resources

Organizing team.

The organizing team comprises a diverse and complementary group of researchers spanning top-tier institutions, including Shenzhen Loop Area Institute (SLAI), The Chinese University of Hong Kong, Sheznzhen, The Chinese University of Hong Kong, The Hong Kong University of Science and Technology, Fudan University, Sun Yat-Sen University, University of Colorado Anschutz, Quebec AI Institute and DexForce. This team exhibits a balanced hierarchy of distinguished professors, assistant professors, and junior researchers (Ph.D., M.S. candidates). While members specializing in robotics and machine learning drive the core technical execution, experts from non-robotics backgrounds provide essential guidance on operational standards, task specifications, and evaluation protocols, aligning directly with all competition assignments:

- **Coordinators:** Guiliang Liu, Weishi Zheng, Kui Jia.
- **Data Providers:** Ruixin Wu, Runyi Zhao, Chengkun Li, Yueci Deng, Hongrui Zhang, Yingying Guo, Lihe Ding, Shaocong Dong, Yanjun Gao, and Yudong Luo.
- **Platform Administrators:** Runyi Zhao and Ruixin Wu.
- **Baseline Method Providers:** Runyi Zhao, Ruixin Wu, Simo Wu, and Tianfan Xue.
- **Beta Testers:** Ruixin Wu and Hongrui Zhang.
- **Evaluators:** Runyi Zhao, Hongrui Zhang, Ruixing Jin, and Ang Li.

Resources provided by organizers. To ensure the successful execution of the RoboSynChallenge, we have secured the following resources: 1) Computing: A cluster of 200 NVIDIA H100 GPUs at SLAI, a cluster of 16 NVIDIA A800 GPUs at CUHKSZ, and 10 GeForce RTX 5090 servers SLAI for model training, dataset generation and leaderboard hosting for participant support. 2) Hardware:

four identical dual-arm workstations featuring AgileX Piper manipulators and RealSense D435i cameras (Figure 2) for physical validation and redundancy. Staff: A 17-person team including PhD researchers, a full-stack engineer, and lab technicians to manage the API, hardware resets, and weekly technical "Office Hours." Sponsors: Chinese University of Hong Kong, Shenzhen Loop Area Institute (SLAI), DexForce, and industry partners providing a 20,000 prize pool and 5,000 in travel grants.

References

[1] Christian Smith, Yiannis Karayiannidis, Lazaros Nalpantidis, Xavi Gratal, Peng Qi, Dimos V. Dimarogonas, and Danica Kragic. Dual arm manipulation - A survey. Robotics and Autonomous Systems, 60(10):1340–1353, 2012.

[2] Oliver Kroemer, Scott Niekum, and George Konidaris. A review of robot learning for manipulation: Challenges, representations, and algorithms. Journal of Machine Learning Research, 22(30):1–82, 2021.

[3] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In Robotics: Science and Systems, RSS, 2023.

[4] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In Robotics: Science and Systems, RSS, 2023.

[5] Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision-language-action models for embodied ai. arXiv preprint arXiv:2405.14093, 2024.

[6] Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. arXiv preprint arXiv:2504.02792, 2025.

[7] Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, et al. Motus: A unified latent action world model. arXiv preprint arXiv:2512.13030, 2025.

[8] Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. arXiv preprint arXiv:2601.21998, 2026.

[9] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. arXiv preprint arXiv:2602.15922, 2026.

[10] Ying Zheng, Lei Yao, Yuejiao Su, Yi Zhang, Yi Wang, Sicheng Zhao, Yiyi Zhang, and Lap-Pui Chau. A survey of embodied learning for object-centric robotic manipulation. Machine Intelligence Research, pages 1–39, 2025.

[11] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. Rlbench: The robot learning benchmark & learning environment. IEEE Robotics and Automation Letters, 5(2):3019–3026, 2020.

[12] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. IEEE Robotics and Automation Letters (RA-L), 7(3):7327–7334, 2022.

[13] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. arXiv preprint arXiv:2306.03310, 2023.

[14] Yao Mu, Tianxing Chen, Shijia Peng, Zanxin Chen, Zeyu Gao, Yude Zou, Lunkai Lin, Zhiqiang Xie, and Ping Luo. Robotwin: Dual-arm robot benchmark with generative digital twins (early version). arXiv preprint arXiv:2409.02920, 2024.

[15] Adina Yakefu, Bin Xie, Chongyang Xu, Enwen Zhang, Erjin Zhou, Fan Jia, Haitao Yang, Haoqiang Fan, Haowei Zhang, Hongyang Peng, et al. Robochallenge: Large-scale real-robot evaluation of embodied policies. arXiv preprint arXiv:2510.17950, 2025.

[16] Yash Jangir, Yidi Zhang, Kashu Yamazaki, Chenyu Zhang, Kuan-Hsun Tu, Tsung-Wei Ke, Lei Ke, Yonatan Bisk, and Katerina Fragkiadaki. Robotarena ∞: Unlimited robot benchmarking via real-to-sim translation. In International Conference on Learning Representations, ICLR, 2026.

[17] Pranav Atreya, Karl Pertsch, Tony Lee, Moo Jin Kim, Arhan Jain, Artur Kuramshin, Clemens Eppner, Cyrus Neary, Edward Hu, Fabio Ramos, et al. Roboarena: Distributed real-world evaluation of generalist robot policies. In Proceedings of the Conference on Robot Learning (CoRL 2025), 2025.

[18] Yiting Chen, Kenneth Kimble, Edward H Adelson, Tamim Asfour, Podshara Chanrungmaneekul, Sachin Chitta, Yash Chitambar, Ziyang Chen, Ken Goldberg, Danica Kragic, et al. Manipulationnet: An infrastructure for benchmarking real-world robot manipulation with physical skill challenges and embodied multimodal reasoning. arXiv preprint arXiv:2603.04363, 2026.

[19] Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations. In Annual Conference on Robot Learning, CoRL, 2023.

[20] Shengliang Deng, Mi Yan, Songlin Wei, Haixin Ma, Yuxin Yang, Jiayi Chen, Zhiqi Zhang, Taoyu Yang, Xuheng Zhang, Wenhao Zhang, et al. Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. arXiv preprint arXiv:2505.03233, 2025.

[21] Guiliang Liu, Yueci Deng, Runyi Zhao, Huayi Zhou, Jian Chen, Jietao Chen, Ruiyan Xu, Yunxin Tai, and Kui Jia. Dexscale: Automating data scaling for sim2real generalizable robot control. In International Conference on Machine Learning, ICML, 2025.

[22] Runyi Zhao, Sheng Xu, Ruixing Jin, Yueci Deng, Yunxin Tai, Kui Jia, and Guiliang Liu. Sim2real VLA: Zero-shot generalization of synthesized skills to realistic manipulation. In International Conference on Learning Representations, ICLR, 2026.

[23] Muhayy ud Din, Waseem Akram, Lyes Saad Saoud, Jan Rosell, and Irfan Hussain. Vision language action models in robotic manipulation: A systematic review. ArXiv, abs/2507.10672, 2025.

[24] Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. arXiv preprint arXiv:2406.02523, 2024.

[25] EmbodiChain Developers. Embodichain: An end-to-end, gpu-accelerated, and modular platform for building generalized embodied intelligence., November 2025.

[26] Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. arXiv preprint arXiv:2511.00062, 2025.

[27] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314, 2025.

[28] Zhengrong Xue, Shuying Deng, Zhenyang Chen, Yixuan Wang, Zhecheng Yuan, and Huazhe Xu. Demogen: Synthetic demonstration generation for data-efficient visuomotor policy learning. arXiv preprint arXiv:2502.16932, 2025.

[29] Tianxing Chen, Kaixuan Wang, Zhaohui Yang, Yuhao Zhang, Zanxin Chen, Baijun Chen, Wanxi Dong, Ziyuan Liu, Dong Chen, Tianshuo Yang, et al. Benchmarking generalizable bimanual manipulation: Robotwin dual-arm collaboration challenge at cvpr 2025 meis workshop. arXiv preprint arXiv:2506.23351, 2025.

[30] Yi-Lin Wei, Haoran Liao, Yuhao Lin, Pengyue Wang, Zhizhao Liang, Guiliang Liu, and Wei-Shi Zheng. Cyclemanip: Enabling cyclic task manipulation via effective historical perception and understanding. arXiv preprint arXiv:2512.01022, 2025.

[31] Ruixiang Wang, Qingming Liu, Yueci Deng, Guiliang Liu, Zhen Liu, and Kui Jia. Eva: Aligning video world models with executable robot actions via inverse dynamics rewards. 2026.

[32] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164, 2024.

[33] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. arXiv preprint arXiv:2504.16054, 2025.

[34] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. RDT-1b: a diffusion foundation model for bimanual manipulation. In International Conference on Learning Representations, ICLR, 2025.

A Technical Details

A.1 Data Collection Protocol in the Real World

To evaluate real-world generalization, each manipulation task was conducted under five experimental conditions (varying background, lighting, and distractors, seen in Table 2), combined with four positions and three orientations. This systematic setup yields 60 diverse samples per task to capture varied spatial and visual factors.

Table 2: Summary of real-world data collection conditions.

Background	Lighting Setting	Additionals	Description
White	Fixed lighting	None	Standard setting with neutral background and lighting.
White	Enhanced lighting	None	Increased illumination to vary lighting conditions.
White	Fixed lighting	2–3 distractors	Includes additional objects to test robustness.
Blue	Fixed lighting	None	Alternative background color to introduce visual contrast.
Yellow	Fixed lighting	None	Textured background used to assess visual generalization.

A.2 Data Generation with EmbodiChain

EmbodiChain [25] is an end-to-end, GPU-accelerated, and modular platform for embodied AI research that integrates high-performance simulation, automated data pipelines, and flexible learning tools, thereby supporting rapid experimentation and effective Sim2Real transfer. It consists of three tightly coupled stages for scalable, diverse, and physically-grounded dataset generation.

1) Generative Simulation of Learning Environment. To overcome the limited diversity of manually designed simulation environments, EmbodiChain employs a two-stage generative framework. First, it synthesizes simulation-ready assets via generative models followed by multi-objective optimization to ensure geometric fidelity, physical plausibility, and simulator compatibility. Second, these assets are composed into fully functional scenes through gradient-based layout synthesis that optimizes object placement for physical realism and robot reachability. This process yields physically consistent, richly annotated environments that form the foundation of large-scale embodied data generation.

2) Data Scaling via Domain Expansion. Building on the generated environments, EmbodiChain scales embodied data by automatically generating and expanding robot interaction trajectories to improve coverage and robustness. It promotes functional diversity through reachability-aware sampling that selects kinematically feasible robot states maximizing task-space dissimilarity (e.g., in end-effector approach direction, contact geometry, and interaction outcomes), reducing trajectory homogenization common in teleoperation and conventional planners. To further strengthen robustness,

a closed-loop error recovery module detects execution failures (e.g., slippage, misaligned grasps, boundary violations) and reactively replans corrective motions, which are relabeled and reintegrated as supervision for recovery behaviors.

3) Sim2Real Generalization via Online Data Streaming. To facilitate scalable Sim2Real transfer, EmbodiChain implements an *Online Data Streaming (ODS)* mechanism that continuously feeds diverse experiences from simulation into the learning loop. A streaming-based visual augmentation module perturbs lighting, textures, and sensor parameters on the fly, enriching perceptual diversity and mitigating overfitting to simulation-specific appearances. In parallel, an asynchronous shared-memory architecture enables real-time data exchange between simulation and learning processes through lock-free circular buffers, maximizing experience throughput and sample efficiency.

Table 3: Summary of simulated data randomization parameters.

Feature	**Parameter**	**Description**
Light	intensity_range: [*min*, *max*]	Lighting intensity variation.
	position_range: [[x_{min}, y_{min}, z_{min}], [x_{max}, y_{max}, z_{max}]]	Light source position randomization.
	color_range: [[0.6, 0.6, 0.6], [1.0, 1.0, 1.0]]	Light color variation.
Object location	init_pos: [x, y, z]	Object initial position.
	position_range: [[$-x$, $-y$, $-z$], [$+x$, $+y$, $+z$]]	Offset range from initial position.
Object orientation	rotation_range: [*min*, *max*]	Initial object rotation randomization.
Object material	base_color_range: [[0.2, 0.2, 0.2], [1.0, 1.0, 1.0]]	Object surface color variability.
Object size	scale: [x, y, z]	Uniform or axis-specific scaling.
Table	position_range: [[0, 0, -0.04], [0, 0, 0.04]]	Table height variation (±4 cm).
	random_texture_prob: p_{texture}	Probability of random table texture.
	base_color_range: [[r_{min}, g_{min}, b_{min}], [r_{max}, g_{max}, b_{max}]]	Table color range.
Background color	base_color_range: [[0.2, 0.2, 0.2], [1.0, 1.0, 1.0]]	Background plate color variation.
Camera (intrinsics)	$f_{x,\text{range}}$	Focal length range along x-axis.
	$f_{y,\text{range}}$	Focal length range along y-axis.
Camera (extrinsics)	pos_range: [[0, -0.02, 0], [0.02, 0.02, 0.01]]	Random camera position (XYZ).
	euler_range: [[-0.175, -0.175, -0.175], [0.175, 0.175, 0.175]]	Camera rotation (roll–pitch–yaw).
Robot init.	eef_pos_range: [[-0.01, -0.01, -0.01], [0.01, 0.01, 0]]	End-effector position variation.
	$q_{\text{pos,range}}$: ([$j_{1,\text{min}}$,...],[$j_{1,\text{max}}$,...])	Joint configuration randomization.
Distractors	Common items (e.g., bowls, cups, toys)	Distractors for scene complexity.

Figure 4: The demonstration showcases our backup hardware, which shares the same robot configurations and setup. This ensures that real-world evaluations can be conducted reliably and at scale.

Table 4: Summary of Task Descriptions, Steps, and Success Criteria

Task Description	Steps	Success Criteria
Click Bell	(1) Move the arm toward the bell. (2) Align and press the button. (3) Return the arm to the initial position.	The button has been successfully pressed within a limited number of action steps.
Items Hand-Over and Place	(1) Grasp the pen. (2) Pass the pen to another arm. (3) Place the pen in the brush pot.	The pen has been successfully placed in the brush pot within a limited number of action steps.
Dual-Arm Water Pouring	(1) Grasp the water bottle and the cup. (2) Pour the water from the water bottle into the cup. (3) Place both objects securely on the table.	The water has been successfully poured into the cup, and both items are placed on the table within a limited number of action steps.
Table Rearrangement	(1) Grasp the spoon and the fork. (2) Arrange them neatly on both sides of the plate.	The spoons and forks have been arranged around the plate within a limited number of action steps.
Basket Pick-and-Place	(1) Grasp the milk box. (2) Place the milk box in the basket. (3) Grasp the basket. (4) Place the basket in the center of the table.	The milk box has been placed into the basket, and the basket has been set at the table center within a limited number of action steps.
Drawer Open and Place	(1) Grasp the handle. (2) Open the drawer. (3) Grasp an item. (4) Place the item in the drawer. (5) Close the drawer.	The item has been successfully placed in the drawer within a limited number of action steps.
Mixer Operating	(1) Grasp the beaker. (2) Place the beaker on the mixer. (3) Start the mixer.	The beaker was placed on the mixer and stirring successfully started within a limited number of action steps.
Item Assembly	(1) Grasp the silicone tubes. (2) Adjust them to a horizontal position. (3) Move one tube forward. (4) Move the other tube and splice together.	The two silicone tubes have been successfully joined together within a limited number of action steps.
Manipulate Pipette	(1) Grasp the pipette. (2) Insert pipette into the first beaker. (3) Another arm presses the pipette button.	The pipette has been correctly inserted into the beaker and the button has been successfully pressed within a limited number of action steps.
Sample Loading	(1) Grasp the test tube. (2) Pass the test tube to another arm. (3) Place the test tube into the rack.	The test tube has been successfully placed in the rack within a limited number of action steps.

B Evaluation Results of baselines

This section presents the baseline evaluation results, which encompass performance comparisons across policies trained on pure simulation data, pure real-world data, as well as the π_0, $\pi_{0.5}$, and Motus model variants. The evaluations are conducted across three core dimensions: success rate, action steps, and inference time.

Crucially, the empirical results show that the metrics of models trained on our simulation data are closely comparable to, and in several scenarios even outperform, those trained on real-world data when deployed in the real world. This demonstrates that our simulation synthesized data has comparable quality with real data, establishing a robust foundation for participants to conduct subsequent Sim2Real policy training and model optimization.

Table 5: Experimental Results of Baselines: Success Rate ($x/20$ or %), Action Steps (max 1000), and Real Time (s).

Model	Click Bell			Items Hand-Over and Place			Dual-Arm Water Pouring			Table Rearrangement		
	SR	Steps	Time	SR	Steps	Time	SR	Steps	Time	SR	Steps	Time
pi0 (sim)	8/20	625.30	**63.78**	5/20	834.75	83.60	6/20	898.60	89.95	7/20	788.20	79.07
pi0 (real)	5/20	860.45	86.05	5/20	844.00	86.10	4/20	917.70	92.69	8/20	742.70	74.27
pi0.5 (sim)	10/20	647.55	65.53	7/20	791.25	**80.00**	7/20	877.45	**87.90**	**12/20**	**612.90**	**61.31**
pi0.5 (real)	6/20	791.20	80.70	5/20	832.90	84.12	6/20	872.15	**87.22**	**12/20**	628.80	63.51
Motus (sim)	**13/20**	**463.30**	79.09	**10/20**	584.70	97.52	**8/20**	**667.30**	111.00	4/20	864.25	143.75
Motus (real)	**14/20**	**420.50**	70.11	**12/20**	**492.30**	83.65	6/20	744.95	125.97	3/20	900.05	153.78

Model	Basket Pick-and-Place			Drawer Open and Place			Mixer Operating			Item Assembly		
	SR	Steps	Time	SR	Steps	Time	SR	Steps	Time	SR	Steps	Time
pi0 (sim)	5/20	835.40	83.56	6/20	821.40	82.23	3/20	897.05	90.09	0/20	1000.00	102.07
pi0 (real)	6/20	796.70	80.47	8/20	757.70	77.29	1/20	967.55	98.69	0/20	1000.00	99.86
pi0.5 (sim)	**10/20**	663.25	**66.42**	11/20	664.40	**67.08**	**4/20**	**864.50**	**87.11**	0/20	1000.00	104.29
pi0.5 (real)	9/20	697.90	71.88	**12/20**	645.70	**65.22**	3/20	901.75	90.18	0/20	1000.00	101.02
Motus (sim)	**10/20**	827.95	132.29	10/20	593.15	102.70	2/20	936.25	154.80	0/20	1000.00	166.22
Motus (real)	9/20	**608.55**	101.27	11/20	**546.60**	93.96	0/20	1000.00	166.41	0/20	1000.00	167.37

Model	Manipulate Pipette			Sample Loading			Task Average		
	SR	Steps	Time	SR	Steps	Time	SR	Steps	Time
pi0 (sim)	2/20	960.65	96.29	2/20	946.85	94.73	22.00%	898.12	90.56
pi0 (real)	0/20	1000.00	108.46	3/20	920.35	92.04	22.50%	881.15	90.20
pi0.5 (sim)	2/20	953.15	**95.82**	**4/20**	**900.05**	**89.97**	**38.50%**	797.55	**80.55**
pi0.5 (real)	**4/20**	**898.75**	**91.67**	3/20	914.95	93.32	33.00%	821.65	82.35
Motus (sim)	**4/20**	905.35	149.33	2/20	945.70	157.82	31.50%	778.80	133.76
Motus (real)	0/20	1000.00	164.84	0/20	1000.00	166.85	27.50%	**721.35**	129.43

C Biography of all team members

- **Runyi Zhao** received the B.S. degree in Physics from Wuhan University, Wuhan, China, in 2024. He is currently working toward the Ph.D. degree in Computer Science with the Chinese University of Hong Kong, Shenzhen, China. His research works have been presented at top-tier venues such as ICLR and ICML. His research interests include embodied intelligence and reinforcement learning.

- **Ruixin Wu** is currently a Ph.D. candidate at the School of Data Science, The Chinese University of Hong Kong, Shenzhen, and a research intern at the Shenzhen Loop Area Institution (SLAI). He received the B.S. degree in Robotics Engineering from Harbin Institute of Technology, Harbin, China, in 2023, and is expected to receive his M.S. degree in Mechanical Engineering from Zhejiang University, Hangzhou, China, in June 2026. His research interests include embodied AI, trajectory planning, and mmWave radar.

- **Hongrui Zhang** is currently pursuing the Master of Philosophy (M.Phil.) degree at the School of Data Science, The Chinese University of Hong Kong, Shenzhen. He is currently an undergraduate student at Harbin Institute of Technology, Weihai, China, expected to receive his Bachelor's degree in June 2026. His research interests include multimodal large language models and embodied intelligence. He has contributed to research published at top-tier venues such as ICLR and IJCAI.

- **Chengkun Li** is currently a Ph.D. candidate at the School of Data Science, The Chinese University of Hong Kong, Shenzhen, and a research intern at the Shenzhen Loop Area Institution (SLAI). He is currently an undergraduate student in Mathematics and Applied Mathematics at Southeast University, China, expected to receive the B.Sc. degree in June 2026. His current research interests include embodied intelligence and reinforcement learning.

- **Ang Li** received the M.S. and B.S. degrees from the School of Mechanical Engineering, Southeast University in 2025 and from the College of Mechanical and Vehicle Engineering, Chongqing University in 2022, respectively. He is currently working toward the Ph.D. degree with the School of Data Science, Chinese University of Hong Kong, Shenzhen and SLAI. His research interests include intelligent perception, embodied AI, and robotics. He serves as a reviewer for TNNLS, TITS, TIM, and IROS.

- **Ruixing Jin** is currently a Ph.D. candidate at The Chinese University of Hong Kong, Shenzhen. He received his Master of Science degree from the University of Michigan and his Bachelor of Engineering degree from Zhejiang University.

- **Yueci Deng** is currently a first-year Computer Science Ph.D. student at the Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen), under the supervision of Prof. Kui Jia. He received his B.S. degree from the University of Electronic Science and Technology of China (UESTC) in 2018, and his M.S. degree from Nanyang Technological University (NTU), Singapore, in 2019. Prior to joining CUHK-Shenzhen, he worked as an architect at DexForce Technology, where he spearheaded the development of Sim2Real AI platform tailored for embodied intelligence.

- **Yingying Guo** is currently pursuing the B.Eng. degree in Computer Science and Engineering at The Chinese University of Hong Kong (Shenzhen), Shenzhen, China, from 2023 to 2027. Her research interests include embodied intelligence and Large Language Model.

- **Tianfan Xue** is an Assistant Professor at the Multimedia Lab (mmlab) in the Department of Information Engineering at the Chinese University of Hong Kong. Prior to this, he worked in the Computational Photography Team at Google Research for over five years. He received his Ph.D. degree from the Computer Science and Artificial Intelligence Laboratory (CSAIL) at the Massachusetts Institute of Technology (MIT) in 2017. He also holds an M.Phil. degree from CUHK, obtained in 2011, and a Bachelor's degree from Tsinghua University. His research focuses on computational photography, 3D reconstruction, and generation. The anti-reflection technology he investigated is utilized by Google Photoscan, which boasts over 10 million users. His recent work on bilateral based 3D reconstruction has won SIGGRAPH Honorable mention 2024, work on HDR fusion won the CVPR Best Demo Honorable Mention. He also served as an area chair for WACV, CVPR, NeurIPS and ACM MM.

- **Lihe Ding** is currently a Ph.D. candidate at MMLab, The Chinese University of Hong Kong, supervised by Prof. Tianfan Xue. He received his B.S. and M.S. degrees from Beijing Institute of Technology. He has worked as a research intern at Kling Gen AI, Kuaishou, and also interned at SenseTime Research. His research interests primarily include 3D generation, world models, and robotics.

- **Shaocong Dong** is currently a Ph.D. candidate in Computer Science and Engineering at The Hong Kong University of Science and Technology, under the supervision of Prof. Dan Xu. He earned his Bachelor's and Master's degrees from Beijing Institute of Technology. Previously, he worked as a research intern at SenseTime Research. His research interests primarily lie in 3D generation, agentic generation, and embodied AI.

- **Yanjun Gao** is an Assistant Professor in the Department of Biomedical Informatics and Co-Director of the Center for Health AI at the University of Colorado Anschutz Medical Campus, where she also leads the LARK Lab. Her research focuses on large language models (LLMs), reinforcement learning, knowledge-guided AI systems, and trustworthy machine learning, with applications spanning healthcare, scientific discovery, and human-AI collaboration. Dr. Gao received her PhD in Computer Science and Engineering from Pennsylvania State University and previously conducted postdoctoral research at the University of Wisconsin–Madison. She is the recipient of an NIH K99/R00 Pathway to Independence Award and has published broadly on LLMs, knowledge graphs, evaluation frameworks, and AI reasoning systems.

- **Yudong Luo** is currently a Postdoctoral Researcher with HEC Montréal and the Mila - Quebec AI Institute, Canada. He received the B.Eng. degree in computer science from Shanghai Jiao Tong University in 2018, the M.Sc. degree from Simon Fraser University in 2020, and the Ph.D.

degree in computer science from the University of Waterloo, Canada, in 2024. He has also been a Visiting Researcher at the Chinese University of Hong Kong, Shenzhen, and the University of Waterloo. His research interests lie at the intersection of reinforcement learning, risk management, and machine learning, with a specific focus on Embodied AI, multi-agent path finding, and sports analytics. He serves as a regular reviewer for top-tier venues such as NeurIPS, ICLR, and ICML, and has published extensively in NeurIPS, ICLR, ICML, and IJCAI.

- **Simo Wu** is currently an Assistant Professor with Fudan University, Shanghai, China. He received the B.S. degree in mathematics from Fudan University and the Ph.D. degree in mathematics from Pennsylvania State University, PA, USA, in 2020. His research focuses on Embodied AI, with an emphasis on video world models for robot manipulation and scaling laws for Transformers. He has authored several publications and technical contributions in these areas, including work on vision-adaptive diffusion policies and nonlinear attention mechanisms for stable model scaling.

- **Kui Jia** received the B.E. degree from Northwestern Polytechnic University, Xi'an, China, in 2001, the M.E. degree from the National University of Singapore, Singapore, in 2004, and the Ph.D. degree in computer science from the Queen Mary University of London, London, U.K., in 2007. He was with the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, Chinese University of Hong Kong, Hong Kong, the Institute of Advanced Studies, University of Illinois at Urbana-Champaign, Champaign, IL, USA, and the University of Macau, Macau, China. He is currently a Professor with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China. His research focuses on theoretical deep learning and its applications in vision and robotic problems, including deep learning of 3D data and deep transfer learning.

- **Wei-shi Zheng** is currently a Full Professor and Director of the Key Laboratory of Machine Intelligence and Advanced Computing (Ministry of Education) at Sun Yat-sen University, China. He received the Ph.D. degree in applied mathematics from Sun Yat-sen University in 2008 and previously served as a Postdoctoral Researcher at Queen Mary University of London, U.K. A Cheung Kong Scholar Distinguished Professor and IAPR Fellow, his research focuses on person re-identification, action prediction, and large-scale machine learning. He has received the NSFC for Excellent Young Scientist and serves as an Associate Editor for IEEE TPAMI and Artificial Intelligence Journal. He has also acted as an Area Chair for NeurIPS, CVPR, and ICCV.

- **Guiliang Liu** is currently an Assistant Professor at the School of Data Science (SDS) in The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen). Prior to joining CUHK-Shenzhen, he worked as a postdoctoral fellow at the University of Waterloo and the Vector Institute with Prof. Pascal Poupart (2020–2022). He received his Ph.D. degree from the School of Computing Science, Simon Fraser University, under the supervision of Prof. Oliver Schulte (2016–2020). He also completed a research internship at the Cognitive Computing Lab, Baidu Research. He received his Bachelor's degree from the School of Computing Science and Engineering, South China University of Technology. His research interests lie in reinforcement learning and embodied AI, with a specific focus on whole-body control for humanoid robots, vision-language-action (VLA) manipulation systems, generative skill acquisition from simulation, and reinforcement learning from human feedback (RLHF). He has served as an Area Chair for top-tier machine learning and AI conferences, including NeurIPS and ICLR.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract and introduction, we mainly demonstrated the importance of coupling synthetic and real data, which is exactly what the proposed competition is trying to encourage the community to solve.

   Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

   Justification: As a competition proposal, we believe that defining specific limitations prior to the competition and the participants' exploration would be premature.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [No]

   Justification: As a competition proposal, this paper focuses less on the theory itself and more on the evaluation protocol and platform construction that are essential for the participants to practice and thus reach conclusions.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The real-world and synthesized data, the data generation platform, together with the checkpoints of all provided baselines, will be fully open-sourced to the participants. Furthermore, a well-defined evaluation protocol ensures high reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The real-world and synthesized data, the data generation platform, together with the checkpoints of all provided baselines, will be fully open-sourced to the participants.

   Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Experimental settings are well introduced in the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper has reported appropriate information about the statistical significance of the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: As a competition proposal, this paper introduces the computational resources needed to train and deploy the baselines, together with the computational resources that will be provided for the participants.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both the potential positive and negative societal impacts of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The paper describes safeguards that have been put in place for responsible release of data or models that have a high risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators or original owners of assets used in the paper are properly credited and the license and terms of use are explicitly mentioned and properly respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: New assets introduced in the paper are well documented, and the documentation is provided alongside the asset.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The paper describes potential risks incurred by study participants.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper describes the usage of LLMs in method of this research.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.