Task-Level Demonstration Data for Vision–Language–Action Models: A Survey of Schemas, Adapters, and Cross-Embodiment Transfer

Masashi
Menily Intelligence
Shenzhen, China
Masashi@Menily.AI

April 2026

Abstract

Training vision–language–action (VLA) models for embodied AI requires *task-level demonstration data*—units that couple a natural-language instruction, a visual context, an action trajectory, and a body morphology specification into a single semantically closed unit. While trajectory-level datasets (Open X-Embodiment, DROID) and motion-level datasets (BONES-SEED, AMASS) have reached a degree of standardization, the task-level semantic layer that sits between them remains fragmented. This fragmentation is the primary barrier to cross-institutional data pooling and cross-embodiment transfer.

We survey the state of task-level demonstration data for VLA training across four dimensions: (1) existing public datasets and their structural assumptions, (2) schema proposals and their controlled vocabularies, (3) heterogeneous-source adapter pipelines for POV video, VR, MoCap, and teleoperation data, and (4) cross-embodiment transfer mechanisms via retargeting. We categorize twelve publicly-announced systems (2023–2026) under a unified taxonomy and identify the structural gap in the task-level semantic layer. We then propose `menily/schema v1`, an open task-level demonstration data specification that defines six top-level fields (`task_id`, `language`, `visual`, `action`, `body`, `meta`) designed to interoperate with Open X-Embodiment / RLDS downstream and NVIDIA SOMA / BONES-SEED upstream. We discuss open problems including long-horizon task decomposition, multi-agent data representation, and whole-body loco-manipulation boundaries.

Keywords: embodied AI; vision–language–action models; demonstration data; schema design; cross-embodiment transfer; retargeting; data specification; robot learning

1 Introduction

Vision–language–action (VLA) models have emerged as the dominant architecture for embodied AI systems that must execute natural-language instructions in physical environments [5, 12, 23]. From Open X-Embodiment's RT-X family [23] to Physical Intelligence's π_0 [3], from NVIDIA GR00T [19] to Google DeepMind's Gemini Robotics [7], VLA models consume a specific form of training data: sequences where a goal-conditioned language prompt is paired with visual observations and an action trajectory executed on a specific embodiment.

The quality, scale, and distribution of this training data increasingly determine downstream policy performance. Physical Intelligence reports that π_0 required over 10,000 hours of robot data for pretraining [3]. OpenVLA scaled to 970k episodes [12]. The recent Ψ_0 model from USC achieved state-of-the-art performance on loco-manipulation benchmarks using 829 hours of

human video combined with only 31 hours of robot data, outperforming baselines trained on 10× the robot data [29].

Despite these advances, the community faces a persistent infrastructure problem: **the data format at the task-level semantic layer is not standardized**. Each laboratory and company defines its own schema—including action space conventions, language annotation protocols, morphology declarations, and segmentation rules. This fragmentation makes inter-institutional data pooling prohibitively expensive, post-hoc conversion costs frequently exceed the cost of re-collection, and cross-embodiment transfer remains a manual effort per target body [2].

This survey contributes four things:

1. A **unified taxonomy** of twelve task-level demonstration data systems, spanning trajectory-level datasets (Open X-Embodiment, DROID, BridgeData V2, OXE-AugE), motion-level datasets (BONES-SEED, AMASS, LAFAN1), and end-to-end pipelines (π_0, OpenVLA, GR00T N1, Gemini Robotics, Ψ_0).
2. A **structural characterization** of the missing middle layer—the task-level semantic interface that connects language goals to executable trajectories.
3. A **proposal**: `menily/schema v1`, an open specification for task-level demonstration data.
4. A discussion of **open problems** for future work.

The remainder of this paper is organized as follows. Section 2 establishes the problem context. Section 3 surveys existing public datasets. Section 4 categorizes schema design patterns. Section 5 reviews cross-embodiment transfer mechanisms. Section 6 presents the `menily/schema v1` proposal. Section 7 discusses open problems. Section 8 concludes.

2 Background and Problem Formulation

2.1 What is "task-level" data?

We define a task-level demonstration as a tuple:

$$\tau = (G_{\text{lang}},\ \mathcal{V},\ \mathcal{A},\ \mathcal{B},\ \mathcal{M})$$

where:

- G_{lang} is a natural-language goal specification, possibly with paraphrases across languages;
- $\mathcal{V}$ is a visual context sequence (frames with camera intrinsics and viewpoint type);
- $\mathcal{A}$ is an action trajectory in a well-defined action space (end-effector 6-DoF, joint N-DoF, or whole-body M-DoF);
- $\mathcal{B}$ is a body morphology specification (kinematic tree, DoF mapping, link lengths);
- $\mathcal{M}$ is metadata (source type, collection region, quality flags, timestamps).

The task-level formulation differs from trajectory-level data (which records (observation, action) pairs without explicit semantic units) and from motion-level data (which captures human body motion without target-task coupling or embodiment binding).

2.2 Why task-level matters for VLA

A VLA model with an action head receives language as a long-horizon conditioning signal, not a per-frame classifier target. For the model to learn the mapping from language goals to coherent action sequences, the training data must expose task boundaries, not merely frame sequences with verb tags [17].

Three failure modes arise when training on non-task-level data:

1. **The action head receives the wrong target**: Frame-level verb annotations train a per-timestep classifier rather than a goal-conditioned trajectory policy.
2. **Task boundaries become lossy post-hoc**: Sliding-window segmentation is systematically miscalibrated for difficult tasks, injecting noise exactly where the training signal is most valuable.
3. **Language distribution shifts**: Per-frame captions are linguistically impoverished (e.g., "hold", "tilt") and do not match the distribution of user instructions at deployment (e.g., "pour water from the blue cup into the kettle") [15].

2.3 The fragmentation problem

We characterize the current state of task-level data as *structurally fragmented*. Specific symptoms:

- **Action space is implicit**: Datasets record actions in capture-robot native format (joint positions, end-effector quaternion+translation, or Euler angles), without declaring the convention.
- **Morphology is undocumented**: Most datasets do not annotate DoF mappings or kinematic tree identifiers, making cross-embodiment use a manual rewrite per target robot.
- **Language is thin**: Single English instruction per episode, no paraphrases, no multilingual variants.
- **Viewpoint is unlabeled**: Ego, third-person, and overhead viewpoints are mixed without explicit labels, degrading visual encoder performance.

These symptoms are not algorithmic problems—they are infrastructure coordination problems. Addressing them requires a common schema, not a new model architecture [14].

3 Survey of Existing Datasets

We organize existing public datasets along two axes: *data stratum* (trajectory / motion / task) and *collection source* (robot teleoperation / human motion / synthetic).

3.1 Trajectory-level datasets

Open X-Embodiment [23] aggregated 60 datasets from 34 laboratories into a unified RLDS format. Scale: 1M+ trajectories, 22 embodiments. This is the de-facto standard for robot-side manipulation data. Its language annotation is limited to single-language instructions per episode.

OXE-AugE [24] extended Open X-Embodiment through synthetic augmentation, reaching 4.4M trajectories (3× original scale). Empirical results show a 24–45% success rate improvement on unseen robot-gripper combinations.

DROID [11] focused on "in-the-wild" manipulation diversity: 76k trajectories, 350 hours, 564 scenes, 86 tasks, 50 collectors. DROID is notable for its explicit coverage of environmental heterogeneity.

BridgeData V2 [31] emphasized task and environment diversity on low-cost manipulation platforms, enabling language- and goal-image-conditioned learning.

3.2 Motion-level datasets

AMASS [13] is the canonical large-scale human motion capture aggregate, containing 45+ hours of mocap data unified under the SMPL body model. Primarily used as a motion prior for humanoid learning.

LAFAN1 [27] is Ubisoft's publicly released motion dataset, widely used in locomotion and general-motion learning research.

BONES-SEED [4], released at GTC 2026 by Bones Studio, marks the first dataset explicitly "purpose-built for humanoid robotics". Scale: 142,220 high-fidelity human motion sequences with language annotations, temporal segmentation, and skeletal metadata. Provided in NVIDIA SOMA and Unitree G1 formats.

PHUMA [6] offers physics-grounded humanoid locomotion data, curated via physics-aware filtering to remove motions infeasible on Unitree G1/H1-2.

3.3 End-to-end VLA systems

Several recent systems combine datasets with foundation models in end-to-end pipelines, making their internal schemas de-facto standards by usage:

- **OpenVLA** [12]: a 7B parameter VLA foundation model pretrained on 970k robot episodes from Open X-Embodiment. Action representation follows RLDS TensorSpec.
- π_0 **/ openpi** [3]: Physical Intelligence's generalist policy, pretrained on 10,000+ hours of multi-robot manipulation data. Internal schema supports six action spaces; training data is proprietary.
- **GR00T N1** [19]: NVIDIA's dual-system humanoid foundation model, combining a vision-language module and a diffusion transformer for real-time motion generation. Training mixture: real robot trajectories, first-person human video, synthetic data.
- Ψ_0 **(Psi-Zero)** [29]: USC's loco-manipulation foundation model using staged training—VLM pretraining on 829 hours of EgoDex human video followed by flow-based action expert post-training on 31 hours of robot data. Fully open-source pipeline.
- **Gemini Robotics** [7]: Google DeepMind's embodied-reasoning model, collaborating with Apptronik humanoid hardware. Training mixture and internal schema not publicly released.
- **ULTRA** [28]: A physics-driven neural retargeting framework validated on Unitree G1, supporting both dense motion reference and sparse task specification.

3.4 Observation: the missing stratum

Across the above twelve systems, a structural pattern emerges. Trajectory-level data has Open X-Embodiment / RLDS as de-facto standard. Motion-level data has BONES-SEED / SOMA emerging as a standard. **The task-level semantic layer has no de-facto standard.** Each end-to-end system (π_0, GR00T, Gemini Robotics) has its own internal schema, but these are not publicly released as interoperable specifications.

4 Taxonomy of Schema Design Patterns

We identify three recurring schema design patterns in existing systems.

4.1 Trajectory-first schemas

Representative: Open X-Embodiment / RLDS, BridgeData V2.

Structure: (observation, action) time-series pairs with episode-level metadata. Language is a secondary annotation on episode headers.

Strengths: Compact, efficient storage; well-suited for trajectory-level policy learning.

Weaknesses: Language is thin; viewpoint and morphology often implicit; cross-embodiment transfer requires manual schema mapping.

4.2 Motion-first schemas

Representative: BONES-SEED, AMASS, LAFAN1.

Structure: Motion clips parameterized by a body model (SMPL, SOMA canonical topology, or similar), with optional language and scene annotations.

Strengths: Unified body representation via SOMA [20] enables cross-format motion aggregation.

Weaknesses: Task semantic layer is absent; retargeting to specific robot embodiments is left to downstream tooling.

4.3 Task-first schemas (emerging)

Representative: Internal schemas of π_0, GR00T, Gemini Robotics; proposed `menily/schema v1` (Section 6).

Structure: A task is a closed semantic unit combining goal language, visual context, action trajectory, body morphology, and metadata. Controlled vocabularies for action space, viewpoint, and source type.

Strengths: Direct match to VLA model input requirements; supports cross-embodiment transfer via explicit morphology annotation.

Weaknesses: No publicly released, interoperable specification exists as of April 2026 (except `menily/schema v1`).

4.4 Schema design decision axes

Across these three patterns, we identify five recurring design decisions:

1. **Language annotation granularity**: per-frame / per-segment / per-episode / per-task.
2. **Action space specification**: implicit (capture-native) / declared / controlled vocabulary.
3. **Morphology declaration**: absent / text description / structured DoF map.
4. **Viewpoint handling**: unlabeled / free text / controlled vocabulary.
5. **Source provenance**: unlabeled / text field / controlled vocabulary.

Different design choices optimize for different downstream uses. Our position in Section 6 is that task-level VLA training requires controlled vocabularies on decisions 2, 3, 4, and 5, and multi-granularity language support on decision 1.

5 Cross-Embodiment Transfer Mechanisms

A central use case for task-level demonstration data is cross-embodiment transfer: a demonstration collected on one body morphology should, at least partially, be usable for training policies on a different body.

5.1 The problem

Given a source demonstration τ_s collected on body B_s and a target body B_t with different DoF layout, link lengths, and actuator characteristics, produce a compatible training signal τ_t such that policies trained on τ_t are viable for B_t [25].

The naive approach—direct trajectory replay—fails due to kinematic and dynamic mismatches. Classical retargeting via inverse kinematics handles kinematic mismatch but ignores dynamic feasibility (contact forces, actuator torque limits, self-collision) [2].

5.2 Recent advances

AdaMorph [1]: A neural retargeting framework that handles zero-shot retargeting across 12 humanoid morphologies. Models retargeting as conditional generation into a morphology-agnostic latent intent space, modulated by embodiment constraints via Adaptive Layer Normalization.

SPARK [9]: Skeleton-parameter alignment combined with three-stage progressive optimization (kinematic TO $\rightarrow$ inverse dynamics $\rightarrow$ full kinodynamic TO), producing dynamically consistent state and torque trajectories.

KDMR [10]: Models retargeting as multi-contact whole-body trajectory optimization, introducing rigid-body dynamics and contact complementarity constraints.

OmniRetarget [22]: Introduces interaction mesh to preserve spatial and contact relationships between robot, terrain, and manipulated objects. Produced 8–9 hours of physically feasible trajectories used in Unitree G1 training.

TWIST2 [32]: Reduces full-body teleoperation collection cost to $\sim$\$250 (PICO4U VR headset plus custom 2-DoF neck), enabling 100 demonstrations in 15 minutes with near-100% success rate.

5.3 Implications for schema design

Effective cross-embodiment transfer requires the source schema to **explicitly declare**:

1. Action space (receives body-relative or world-relative semantics).
2. Morphology identifier and DoF map.
3. Link lengths (for length-aware retargeting).
4. Invariant landmarks—task-relevant keyframes that any target embodiment must reach.

The fifth requirement (invariant landmarks) is the key design choice: it transforms a rigid per-robot trajectory into a **waypoint schema with continuous interpolation**, which is the representation cross-embodiment transfer actually requires [14].

6 The `menily/schema v1` Specification

Building on the taxonomy of Sections 3–5, we propose `menily/schema v1` as a candidate task-level demonstration data specification. The full specification is available at `https://github.com/MenilyIntelligence/schema`.

6.1 Design principles

1. **Task-first**: A demonstration is a closed semantic unit, not a time-series with side annotations.
2. **Controlled vocabularies for interoperability**: Action space, viewpoint, morphology, and source type are enumerated types, not free text.

3. **Explicit body annotation for cross-embodiment**: `morphology` and `dof_map` are required fields.

4. **Multi-language support by design**: `language.variants` is recommended-required, not optional.

5. **Scope discipline**: The schema deliberately excludes reward fields (not an RL data format), scene graphs (downstream parsing), and human biometrics.

6.2 Schema definition

A task-level demonstration in `menily/schema v1` is a JSON object with six top-level fields. An abbreviated example:

```
{
  "schema_version": "menily.task-demo/1",
  "task_id": "uuid",
  "language": {
    "instruction": "Pour water from the blue cup into the kettle.",
    "language_code": "en",
    "variants": ["..."]
  },
  "visual": {
    "frames": "path/to/frames/",
    "fps": 30,
    "camera_intrinsics": { "fx": 1128.5, "fy": 1128.5,
                           "cx": 960, "cy": 540 },
    "viewpoint": "ego"
  },
  "action": {
    "space": "ee_6dof",
    "trajectory": [ /* N x action_dim */ ],
    "timestamps": [ /* N */ ],
    "gripper": [ /* N x 1 */ ]
  },
  "body": {
    "morphology": "bimanual_humanoid",
    "dof_map": {
      "right_arm": [0,1,2,3,4,5,6],
      "left_arm":  [7,8,9,10,11,12,13]
    }
  },
  "meta": {
    "source": "pov_video",
    "collection_region": "SEA",
    "collection_time": "2026-01-14T08:20:00Z",
    "quality_flags": ["no_slip","no_contact_gap"]
  }
}
```

The `language.variants` field supports multilingual paraphrases, e.g., the Chinese equivalent of the above English instruction: 把蓝色杯子里的水倒进水壶里.

6.3 Field rationale

Table 1 summarizes each field's design decision and empirical motivation.

6.4 Interoperability

`menily/schema v1` is designed to interoperate with existing standards:

Field	Design decision	Rationale
`language.variants`	Recommended-required	Multi-language paraphrase is near-zero marginal cost (LLM-generated) and critical for deployment robustness [15].
`visual.viewpoint`	Controlled vocabulary	Ego/third-person/overhead signals differ qualitatively for visual encoders [18].
`action.space`	Controlled vocabulary	Implicit action spaces are the most common cross-dataset incompatibility [23].
`body.morphology`	Required	Cross-embodiment transfer infeasible without explicit morphology [1].
`body.dof_map`	Required	DoF index-to-joint mapping is non-discoverable from raw trajectories.
`meta.source`	Controlled vocabulary	Different sources have qualitatively different noise characteristics [21].
`meta.collection_region`	Top-level field	Geographic distribution analysis is critical for bias auditing [16].

Table 1: Design decisions and rationale for `menily/schema v1` fields.

- **Downstream to RLDS**: `Task.to_rlds()` exports an RLDS-compatible episode bundle, preserving the schema's semantic annotations in episode metadata.
- **Downstream to HuggingFace Datasets**: `Task.to_hf_dataset()` produces a `datasets.Dataset` object for direct consumption by HF-based training pipelines.
- **Upstream from BONES-SEED**: `body.morphology` and `body.dof_map` align with SOMA canonical topology, supporting direct consumption of BONES-SEED motion data with task-level semantic overlay.
- **Bidirectional with RLDS**: `from_rlds()` enables conversion of existing Open X-Embodiment / RLDS datasets into `menily/schema` format, augmenting them with task-level semantic information.

6.5 Out-of-scope items

Items explicitly excluded from v1:

- **Reward / return-to-go fields**: `menily/schema` is not an RL data format. Reinforcement learning data should use D4RL or RLDS.
- **Complete scene graphs**: Visual tokens are derived from frames; scene parsing is downstream.
- **Human biometric metadata**: Not collected, no schema field reserved.
- **Embedded URDF/MJCF**: Body morphology is represented as a compact index; full physics simulation models are referenced externally.

6.6 The `menily/toolkit` reference implementation

Accompanying the schema is `menily/toolkit`, an Apache-2.0 Python library providing three adapters:

- `toolkit.pov`: First-person video → task-level data (hand keypoint detection, trajectory reconstruction, task segmentation).

- `toolkit.vr`: Quest / Vision Pro / PICO hand-tracking → end-effector trajectories.
- `toolkit.mocap`: BVH / FBX motion capture → full-body action sequences, with integrated AdaMorph, OmniRetarget, and SPARK retargeting backends.

The toolkit is available at https://github.com/MenilyIntelligence/toolkit.

7 Open Problems

The `menily/schema v1` specification addresses the core fragmentation problem but leaves several open problems for future work.

Long-horizon task decomposition. Tabletop manipulation admits clean task boundaries (grasp, transport, release). Long-horizon tasks (preparing a meal, tidying a room) do not—they decompose into sub-tasks whose boundaries are non-unique. A hierarchical schema with explicit sub-task relationships is an open design question.

Multi-agent scenarios. Two or more robots cooperating on a single task requires a multi-agent task representation. Candidate approaches include: (a) per-agent task files with shared `task_id`, (b) a single file with per-agent action tracks. Neither has been adopted in existing specifications. v1 supports only single-agent demonstrations.

Whole-body loco-manipulation boundaries. The semantic boundary of a loco-manipulation task (walking while carrying while manipulating) is difficult to define cleanly. Existing formulations either treat the whole activity as a single task (losing granularity) or split into locomotion + manipulation sub-tasks (losing the coupling). A principled decomposition remains open.

Quality metrics for data ingestion. Schema conformance is a necessary but insufficient quality criterion. Downstream training quality also depends on retargeting artifact levels, physical feasibility of trajectories, contact-timing consistency, and task-language alignment. Standardizing quality metrics—perhaps as a scoring function over `meta.quality_flags`—is an open research direction [26].

The role of synthetic data. As synthetic augmentation [24] and sim-to-real pipelines [8, 30] mature, schemas need explicit provenance for synthetic vs. real data. v1's `meta.source` includes `sim_generated` as a value, but richer provenance (simulator identity, physics fidelity, domain randomization parameters) may be needed.

Governance and standardization. Schemas that succeed as de-facto standards typically do so through sustained community adoption rather than top-down declaration. Open X-Embodiment's success is partly because it was an inclusive aggregation effort. Whether `menily/schema v1` achieves similar adoption is an empirical question that will play out over the coming 12–24 months. We believe the task-level layer is currently in a standardization window; we invite critique and contribution.

8 Conclusion

The task-level semantic layer of embodied AI training data remains structurally fragmented as of April 2026, despite standardization successes at the trajectory level (Open X-Embodiment, RLDS) and motion level (BONES-SEED, SOMA). This fragmentation is the primary barrier to cross-institutional data pooling, cross-embodiment transfer, and multi-lingual VLA training.

We surveyed twelve task-level demonstration data systems spanning 2023–2026 and identified three recurring schema design patterns (trajectory-first, motion-first, task-first). We proposed `menily/schema v1`, an open task-level specification with controlled vocabularies for action space, viewpoint, morphology, and data source, designed to interoperate with both Open X-Embodiment / RLDS downstream and BONES-SEED / SOMA upstream.

The schema is a draft, not a final standard. We invite community critique via GitHub Issues at https://github.com/MenilyIntelligence/schema and direct technical discussion via email. Companion tooling (`menily/toolkit`) is in internal alpha, with PyPI release planned for the coming weeks.

The infrastructure problem of embodied AI data is solvable in principle. It requires not a new model architecture but a sustained community convergence on a shared format. We believe the coming 12–24 months are the window in which that convergence either happens or fails.

Acknowledgments

The author thanks the Menily Intelligence team for their work on `menily/schema` and `menily/toolkit`, and the broader embodied AI community whose public datasets and tooling made this survey possible. This work benefited from informal discussions with researchers at multiple VLA laboratories. All errors are the author's own.

Author Statement

This paper is authored by Masashi, founder of Menily Intelligence, a company headquartered in Shenzhen that builds task-level demonstration data infrastructure for VLA and humanoid robotics teams. The proposed `menily/schema v1` is the author's own work.

To mitigate reviewer concerns about self-promotion: Sections 2–5 survey existing systems based solely on public information, with no preferential treatment for Menily's own approach; Section 6 is clearly labeled as the author's own contribution; all cited works are verifiable via their listed URLs and arXiv IDs; open problems (Section 7) honestly enumerate limitations of the v1 proposal.

Correspondence: `Masashi@Menily.AI`. Code and specification: https://github.com/MenilyIntelligence
Organization website: https://menily.ai.

References

[1] AdaMorph Team. AdaMorph: Cross-morphology zero-shot retargeting. *arXiv preprint arXiv:2601.07284*, 2026. URL https://arxiv.org/abs/2601.07284.

[2] João Araújo et al. Retargeting matters: Quantifying the impact of motion retargeting on humanoid policy robustness. https://jaraujo98.github.io/retargeting_matters/, 2025.

[3] Kevin Black et al. π_0: A vision-language-action flow model for general robot control. Physical Intelligence, 2024. URL https://www.pi.website/blog/pi0.

[4] Bones Studio. BONES-SEED: The first multimodal motion dataset purpose-built for humanoid robotics. GTC 2026, https://huggingface.co/datasets/bones-studio/seed, 2026.

[5] Anthony Brohan et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. URL https://arxiv.org/abs/2307.15818.

[6] DAVIAN Robotics. PHUMA: Physics-grounded humanoid locomotion dataset. https://github.com/DAVIAN-Robotics/PHUMA, 2025.

[7] Google DeepMind. Gemini robotics: Bringing AI into the physical world. https://deepmind.google/blog/gemini-robotics-brings-ai-into-the-physical-world/, 2025.

[8] HALO Team. HALO: Differentiable simulation for heavy-load agile humanoid motion. *arXiv preprint arXiv:2603.15084*, 2026. URL https://arxiv.org/abs/2603.15084.

[9] Intelligent Control Lab. SPARK: Skeleton-parameter aligned kinodynamic retargeting. *arXiv preprint arXiv:2603.11480*, 2026. URL https://arxiv.org/abs/2603.11480.

[10] KDMR Team. Kinodynamic multi-contact retargeting via trajectory optimization. *arXiv preprint arXiv:2603.09956*, 2026. URL https://arxiv.org/abs/2603.09956.

[11] Alexander Khazatsky et al. DROID: A large-scale in-the-wild robot manipulation dataset. https://droid-dataset.github.io, 2024.

[12] Moo Jin Kim et al. OpenVLA: An open-source vision-language-action model. https://openvla.github.io, 2024. Stanford, UC Berkeley, TRI, Google DeepMind.

[13] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. https://amass.is.tue.mpg.de, 2019.

[14] Menily Intelligence. Cross-embodiment transfer in task-level demonstration data. Research note, https://github.com/MenilyIntelligence/research, 2026.

[15] Menily Intelligence. The data gap in embodied AI, stated precisely. Research note, https://github.com/MenilyIntelligence/research, 2026.

[16] Menily Intelligence. Regional distribution auditing in embodied AI training data. Working note, 2026.

[17] Menily Intelligence. Task-level abstraction: Why frame-level annotation breaks VLA. Research note, https://github.com/MenilyIntelligence/research, 2026.

[18] Suraj Nair et al. R3M: A universal visual representation for robot manipulation. In *CoRL 2022*, 2022.

[19] NVIDIA GR00T Team. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025. URL https://arxiv.org/abs/2503.14734.

[20] NVIDIA Labs. SOMA / SOMA-X: A unified parametric human body representation. *arXiv preprint arXiv:2603.16858*, 2026. URL https://arxiv.org/abs/2603.16858.

[21] NVIDIA Research. What matters in learning from large-scale datasets for robot manipulation. *arXiv preprint arXiv:2506.13536*, 2025. URL https://arxiv.org/abs/2506.13536.

[22] OmniRetarget Team. OmniRetarget: Interaction-preserving data generation for humanoid learning. In *ICRA 2026*, 2026.

[23] Open X-Embodiment Collaboration. Open X-Embodiment: Robotic learning datasets and RT-X models. *arXiv preprint arXiv:2310.08864*, 2023. URL https://arxiv.org/abs/2310.08864.

[24] OXE-AugE Team. OXE-AugE: Large-scale augmented robot manipulation dataset. *arXiv preprint arXiv:2512.13100*, 2025. URL https://arxiv.org/abs/2512.13100.

[25] Xue Bin Peng et al. Learning agile robotic locomotion skills by imitating animals. In *Robotics: Science and Systems (RSS)*, 2020.

[26] Carmelo Sferrazza et al. HumanoidBench: Simulated humanoid benchmark for whole-body locomotion and manipulation. https://humanoid-bench.github.io, 2024.

[27] Ubisoft La Forge. LA-FORGE animation dataset (LAFAN1). https://github.com/ubisoft/ubisoft-laforge-animation-dataset, 2020.

[28] ULTRA Team. ULTRA: Unified multi-modal loco-manipulation on Unitree G1. *arXiv preprint arXiv:2603.03279*, 2026. URL https://arxiv.org/abs/2603.03279.

[29] USC Physical Superintelligence Lab. Ψ_0 (Psi-Zero): A staged-training foundation model for humanoid loco-manipulation. https://psi-lab.ai/Psi0, 2026.

[30] VIRAL Team. VIRAL: Sim-to-real loco-manipulation via teacher-student distillation. *arXiv preprint arXiv:2511.15200*, 2025. URL https://arxiv.org/abs/2511.15200.

[31] Homer Walke et al. BridgeData V2: A dataset for robot learning at scale. RAIL Lab, UC Berkeley, https://rail-berkeley.github.io/bridgedata/, 2023.

[32] Yanjie Ze et al. TWIST2: Portable, full-stack humanoid data collection. In *ICRA 2026*, 2026.