AoE: Always-on Egocentric Human Video Collection for Embodied AI

Supplementary Material

7. Appendix

7.1. Detailed Rating Criteria

This section details the metrics and evaluation standards used to compare data collection paradigms in Table 1. We assess each method based on its potential to scale toward the "foundation model" era of robotics.

Cost (per user): This metric represents the total financial investment required to equip a single participant for high-quality data collection. **Teleoperation** involves the highest barrier ($>$ \$50k) due to industrial arms and specialized master–slave setups. **UMIs** (\$300–\$800) require GoPro cameras and customized 3D-printed hardware. **Wearables** ($>$ \$2k) typically involve high-end VR headsets or motion-capture gloves. **Passive Videos** are considered free as they leverage existing internet resources. **AOE** ($<$ \$20) achieves ultra-low cost by utilizing a user's existing smartphone or action camera paired with a minimal, low-cost mounting solution.

Non-Intrusiveness: This measures how naturally a user can perform tasks without being hindered by hardware. **Teleoperation** (★☆☆☆☆) is highly intrusive, often requiring the user to be tethered to a specific station. **UMIs** (★★★☆☆) require the user to hold a specific tool instead of using their hands directly. **Wearables** (★☆☆☆☆–★★☆☆☆) often involve heavy headsets or restrictive suits. **Passive Videos** (★★★★★) represent natural daily life. ***AoE*** (★★★★☆) allows for hands-free, natural dexterous manipulation, though the presence of a lightweight mount prevents a perfect score.

Scalability: This evaluates the feasibility of expanding the data collection pool to thousands of non-expert users globally. Methods requiring expensive hardware (**Teleoperation**, **Wearables**) have low scalability (★☆☆☆☆–★★☆☆☆). **UMIs** (★★★☆☆) are moderately scalable but still require shipping hardware. **Passive Videos** and ***AoE*** (★★★★★) are the most scalable because they rely on ubiquitous consumer electronics (smartphones) or existing web data.

Deployment Ease: This reflects the technical friction involved in setting up the system. **Teleoperation** and **UMIs** (★☆☆☆☆–★★☆☆☆) require complex calibration, 3D printing, or specific mechanical assembly. ***AoE*** (★★★★★) is designed as a "plug-and-play" solution, requiring only a simple attachment and an app, comparable to the ease of uploading **Passive Videos**.

Data Quality: This assesses the utility of the data for training Vision-Language-Action (VLA) models, particularly for learning fine-grained manipulation. **Teleoperation** (★★★★★) provides "gold-standard" direct action labels. **Passive Videos** (★☆☆☆☆) suffer from severe noise, motion blur, and a lack of consistent viewpoints or action labels. ***AoE*** (★★★★☆) provides high-fidelity egocentric views with consistent hand–object interaction dynamics, providing much cleaner signals for policy learning than raw passive video.

7.2. Always-On Collection App Workflow

The AOE application workflow, as illustrated in Fig. 9, is designed to balance autonomous data acquisition with user-controlled privacy. The process begins with explicit camera authorization, enabling the app to enter a continuous monitoring state. In this mode, on-device lightweight models scan for hand-object interactions, automatically triggering high-quality recording when relevant actions are detected. To optimize storage and data relevance, the system autonomously discards clips with insufficient duration while locally saving successful captures that meet the quality threshold.

Following the collection phase, the app provides a suite of management tools for data curation. Users can browse the local gallery and utilize a built-in editor to trim irrelevant or sensitive frames from the recordings. To ensure privacy and data integrity, the app requires users to manually select and validate clips before batch-uploading them to the cloud along with synchronized sensor metadata. Finally, the system updates a personal dashboard, providing statistical feedback on total collection hours and effective interaction time to help users monitor their data contribution efficiency.

7.3. Distributed System Implementation

Large-scale robot learning from human demonstrations requires processing massive volumes of egocentric video data collected from geographically distributed edge devices. To support this objective, we designed a distributed edge-cloud collaborative architecture that addresses three critical system-level challenges:

System Design Challenges. Existing centralized data collection systems face fundamental limitations when scaling to global deployment:

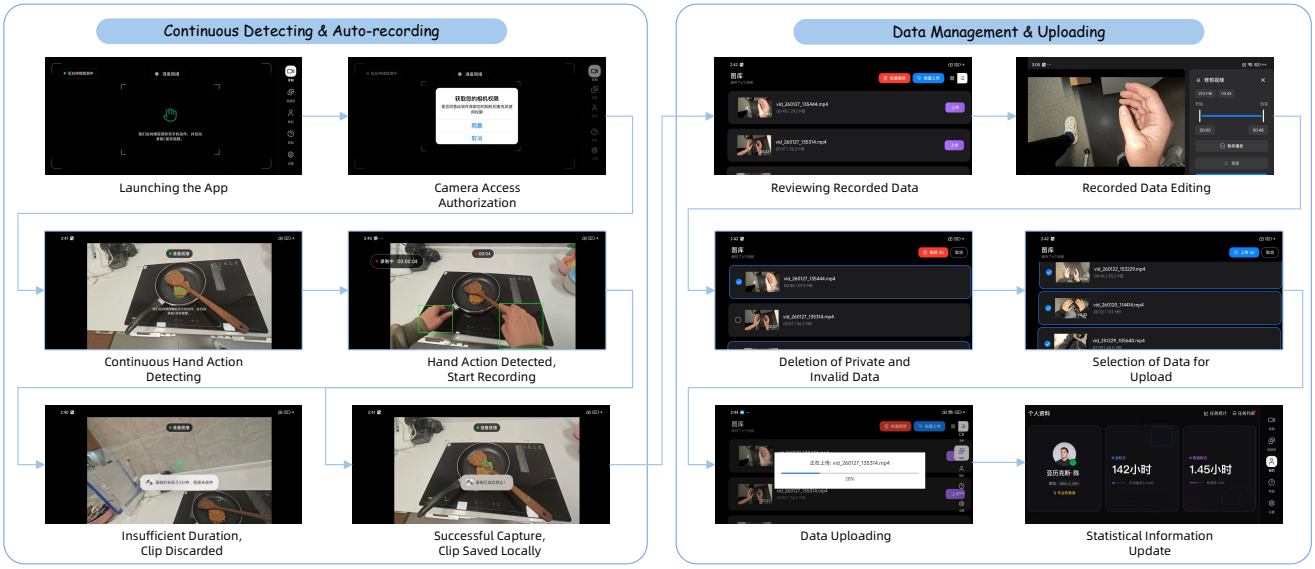

Figure 9. Always-On Collection App Workflow.

- **High-Latency Data Transfer**: Long-distance uploads from geographically dispersed collectors introduce significant network latency (often exceeding 500ms round-trip time) and bandwidth bottlenecks, degrading user experience and limiting upload throughput to impractical levels for high-resolution video streams.
- **Inflexible Processing Pipelines**: Hardcoded processing workflows cannot accommodate rapidly evolving egocentric video analysis algorithms. Integrating new algorithms typically requires weeks of system re-engineering, creating a critical bottleneck for algorithmic innovation.
- **Limited Scalability**: Centralized architectures lack the ability to dynamically allocate computational resources based on real-time workload, resulting in either resource waste during idle periods or processing delays during peak demands. This limitation severely constrains concurrent data streaming from thousands of devices.

To overcome these challenges, our distributed system embodies a decoupled design philosophy across three primary dimensions:

- **Geographic Decoupling**: Proximity-based edge ingestion nodes minimize data transfer latency while maintaining centralized data management through asynchronous synchronization.
- **Algorithmic Decoupling**: Declarative pipeline orchestration enables rapid integration of new processing algorithms without system-level modifications.
- **Resource Decoupling**: Cloud-native elastic scaling dynamically provisions computational resources based on real-time demand, ensuring efficient utilization across varying workloads.

The following sections detail our solutions to each challenge: edge-cloud architecture for low-latency data transfer, customizable processing pipeline for algorithmic flexibility, and elastic scaling mechanisms for dynamic resource management.

7.3.1. Edge-Cloud Collaborative Architecture

Our system adopts a two-tier architecture that balances local access efficiency with centralized processing power, as illustrated in Figure 4.

Distributed Proximity-Based Edge Ingestion. To minimize data transfer latency, we deploy edge ingestion nodes across multiple geographic regions. The system architecture includes:

- **Intelligent Routing**: Collector devices—including smartphones and wearable devices—automatically route to the nearest ingestion point through DNS-based geographic routing and latency-aware selection algorithms.
- **Asynchronous Cross-Region Synchronization**: Data collected at each edge node is aggregated to centralized cloud storage through an hourly asynchronous replication mechanism. This design ensures unified data management while maintaining local access efficiency and tolerating temporary network partitions.

Achieved Benefits. Compared to traditional centralized architectures, our edge-cloud design delivers measurable improvements:

- **Significant Latency Reduction**: Proximity-based edge ingestion reduces average upload latency from 500ms+ to under 100ms, significantly improving user experience and enabling seamless "always-on" data collection.
- **High-Concurrency Device Support**: The distributed architecture supports high-concurrency data streaming from thousands of devices simultaneously, enabling truly global-scale data collection.

7.3.2. Customizable Data Processing Pipeline

To address the challenge of **Inflexible Processing Pipelines**, we designed a highly flexible pipeline orchestration framework. Egocentric video data processing for robot learning is a rapidly evolving field where optimal algorithms and processing paradigms remain undefined. Existing systems typically hardcode specific processing workflows, forcing researchers to either accept suboptimal methods or invest weeks in system re-engineering for each new algorithm.

Custom Operators and Flexible Orchestration. Our system supports user-defined data processing operators and flexible workflow orchestration through declarative configurations, as illustrated in Figure 3. The framework provides three key capabilities:

- **Modular Algorithm Integration**: Researchers can rapidly compose different vision algorithm modules (such as hand pose estimation, object detection and tracking, depth reconstruction, action segmentation, etc.) to build end-to-end data processing pipelines according to experimental needs. Each algorithm is encapsulated as an independent operator with standardized input/output interfaces.
- **Declarative Workflow Configuration**: Users define custom processing workflows, specifying operator sequences, resource requirements, and data dependencies. This declarative approach eliminates the need for low-level system programming and enables rapid pipeline reconfiguration.
- **Hot-Swappable Components**: The system supports runtime replacement of individual operators without pipeline restart, enabling A/B testing of different algorithm versions and seamless algorithm upgrades.

Achieved Benefits. This programmability reduces the integration and validation cycle for new algorithms from traditional weeks to days, greatly accelerating algorithm iteration speed.

7.3.3. Elastic Scaling and Resource Management

To address the challenge of **Limited Scalability**, we designed a cloud-native elastic scaling architecture that dynamically provisions computational resources based on real-time demand. Traditional centralized systems allocate fixed resources, leading to either resource waste during idle periods or processing delays during peak demands—a critical limitation when handling concurrent data streams from thousands of globally distributed devices.

Dynamic Resource Allocation. The system employs a cloud-native resource scheduling architecture built on Kubernetes with the following capabilities:

- **Horizontal Pod Autoscaling (HPA)**: The system monitors queue depth and processing latency metrics in real-time, automatically scaling worker nodes based on demand.
- **Intelligent Resource Partitioning**: GPU-intensive operators (e.g., depth estimation, hand reconstruction) are scheduled on GPU nodes, while CPU-bound operators (e.g., data I/O, format conversion) run on cost-effective CPU nodes. This heterogeneous scheduling optimizes both performance and cost.

Achieved Benefits. Our elastic scaling architecture delivers:

- **High Resource Utilization**: Dynamic scaling maintains high resource utilization efficiency across varying workloads, avoiding both over-provisioning waste and under-provisioning bottlenecks.
- **Rapid Response Time**: The system responds to workload spikes within minutes, enabling seamless transition from pilot studies to large-scale production without manual capacity planning.

7.4. Experiment Setups

We employ FLARE [48] to enable robotic policies to learn from human videos without action labels, such as ***AoE*** data. Unlike teleoperation-based learning paradigms, FLARE [48] is tasked with aligning latent representations of future human actions, ensuring predicted feature embeddings closely match the actual feature vectors of future video frames. This enables the model to learn the physical laws and task logic governing transitions from current to future states by observing vast amounts of human video, even without explicit action labels. In our model design, we transfer this understanding of environmental dynamics to the robot's policy learning.

We fine-tune a pre-trained GR00T N1.5 [4] using the FLARE [48] loss. When training jointly with teleoperation and ***AoE*** data, we apply supervision via future latent alignment loss and action flow matching loss; when using teleoperation data alone, we apply supervision solely via action flow matching loss. During training, we employed fine-tuning steps that scaled positively with dataset size and task difficulty to prevent overfitting. For instance, for the scarf-folding task, we tested models fine-tuned for 100,000 steps on the largest dataset (50 Teleop + 200 ***AoE***). Conversely, for simpler tasks like closing a laptop, we employed only 60,000 fine-tuning steps for the same dataset size.

7.5. User Consent Forms and Privacy Protocols

Overview

This Privacy Policy Summary outlines how ***AoE*** collects, uses, stores, and protects your personal information when you use our platform services, which include video data collection through first-person perspective hand movement recording.

Key Definitions

- **Personal Information:** Any information recorded electronically or otherwise that relates to an identified or identifiable natural person, excluding anonymized data.
- **Video Data:** Video files you record and upload through our platform, documenting hand movements from a first-person perspective.

Information We Collect

1. Account and Registration Information
- **Required Information:** Real name, social media account details (with real-name verification)
- **Purpose:** User identity verification and payment settlement

2. Service Usage Information
- **Video Data:** Hand movement videos you record and upload
- **Technical Data:** Device information, usage logs, platform interactions
- **Payment Information:** Identity and payment details for fee settlement

3. Digital Certificate Information (if applicable)
- **Basic Data:** Age, demographic information
- **Certificate Data:** Academic, driver's license, or vehicle license information
- **Verification Data:** Identity verification results and supporting documents

4. Platform Interaction Data
- **User Content:** Information you voluntarily share or input
- **Communication Data:** Customer service interactions, feedback
- **Technical Logs:** System access records, error reports

How We Use Your Information

1. Service Provision and Management
- **Account Management:** User authentication, account security
- **Service Delivery:** Video data processing, review, and payment
- **Technical Support:** Platform maintenance and troubleshooting

2. Video Data Processing
- **Content Review:** Assessment against usability standards
- **Quality Control:** Technical quality verification
- **Payment Processing:** Fee calculation and disbursement

3. Platform Improvement
- **User Experience:** Service optimization and personalization
- **Feature Development:** New functionality research and implementation
- **Security Enhancement:** Fraud prevention and system protection

4. Legal and Compliance
- **Regulatory Requirements:** Compliance with applicable laws
- **Dispute Resolution:** Investigation and resolution of user issues
- **Legal Obligations:** Response to legal requests and proceedings

Information Sharing and Disclosure

1. Third-Party Service Providers
- **Payment Processing:** Financial service partners for fee settlement
- **Technical Services:** Infrastructure and support providers
- **Business Partners:** Collaborative service delivery partners

2. Legal Requirements
- **Government Authorities:** When required by law or regulation
- **Legal Proceedings:** In response to valid legal requests
- **Protection of Rights:** To protect our legal rights and property

3. Business Transfers
- **Mergers and Acquisitions:** As part of business restructuring
- **Asset Transfers:** During sale or transfer of business assets

Your Rights and Choices

1. Access and Control
- **Account Settings:** Manage your profile and preferences
- **Information Access:** Review your personal data
- **Opt-Out Choices:** Control marketing communications

2. Privacy Preferences
- **Permissions Management:** Control access to device features
- **Data Sharing:** Manage third-party data sharing
- **Communication Preferences:** Choose how we contact you

3. Account Management
- **Profile Updates:** Modify your account information
- **Account Deletion:** Request account termination
- **Data Portability:** Request your data in accessible format

Data Security and Protection

1. Security Measures
- **Technical Safeguards:** Encryption, access controls, firewalls
- **Organizational Measures:** Staff training, security policies
- **Procedural Controls:** Regular security assessments and audits

2. Data Retention
- **Service Data:** Retained for as long as your account is active
- **Legal Requirements:** Retained to comply with legal obligations
- **Business Needs:** Retained for legitimate business purposes

3. Incident Response
- **Breach Notification:** Prompt notification of security incidents
- **Response Procedures:** Established incident management protocols
- **User Protection:** Measures to protect affected users

Important Privacy Considerations

1. Video Data Processing
- **Content Restrictions:** Avoid including personal information in videos
- **Rights Transfer:** Complete intellectual property rights transfer upon upload
- **Usage Rights:** Broad platform rights to use uploaded content

2. Cross-Border Data Transfers
- **Global Operations:** Data may be processed in multiple jurisdictions
- **International Standards:** Compliance with applicable international standards
- **User Awareness:** Acknowledge cross-border data processing implications

Platform-Specific Information

1. Device Requirements
- **Collection Device:** Wearable phone holder provided by platform
- **Personal Device:** Self-provided smartphone with technical specifications
- **Device Security:** User responsibility for proper use and protection

2. Video Standards and Requirements
- **Content Guidelines:** Production-value hand movements only
- **Technical Specifications:** Minimum quality and duration requirements
- **Usage Restrictions:** Prohibited content types and activities

3. Fee Settlement and Payments
- **Eligibility Requirements:** Minimum weekly video duration threshold
- **Payment Processing:** Third-party payment service providers
- **Tax Responsibility:** User obligation for tax declaration and payment

Updates and Changes

1. Policy Updates
- **Modification Process:** Periodic review and updates
- **Notification Methods:** Platform announcements and direct notifications
- **User Acceptance:** Continued use constitutes acceptance of changes

2. Effective Dates
- **Current Policy:** Applies from specified effective date
- **Future Changes:** Take effect according to specified timelines
- **Historical Versions:** Previous versions available upon request

Contact and Support

1. Privacy Inquiries
- **Contact Methods:** Official hotline and support channels
- **Response Times:** Reasonable timeframe for inquiries
- **Complaint Resolution:** Process for addressing privacy concerns

2. Technical Support
- **Device Assistance:** Help with collection device usage
- **Platform Issues:** Technical problem resolution
- **Account Support:** Assistance with account management

3. Legal and Compliance
- **Regulatory Questions:** Information about legal obligations
- **Data Protection:** Details about privacy rights and protections
- **Contractual Terms:** Clarification of user agreement provisions

Important Notice: This summary provides an overview of key privacy considerations but does not replace the full User Agreement. Please read the complete agreement for detailed terms and conditions. By using our services, you

acknowledge and agree to our data processing practices as described in our full privacy policy and user agreement.

Effective Date: As specified in the complete User Agreement *Last Updated:* April 8, 2026

Your continued use of *AoE* services constitutes acceptance of our privacy practices and terms of service.