Deccan AI

DECCAN AI - PHYSICAL INTELLIGENCE RESEARCH

Project Bumblebee

A Curated Egocentric Video Corpus for Frontier Robotics

Animesh Maheshwari Narendar Sriram Nikhil Parlapalli Ankit Khedia Navaneeth Bodla

robotics@deccan.ai

ABSTRACT

We introduce Bumblebee, a multi-modal dataset of synchronized egocentric video, three-camera (head + dual-wrist) body-mounted video, and marker-based motion capture spanning household and factory environments, assembled for training embodied policies, world models, and manipulation systems. The egocentric video stream comprises **11,778** curated hours across **29,247** clips from household settings recorded by **1,802** participants, and **2,544** hours across **1,283** clips from factory settings recorded by **482** participants. This is complemented by **75** hours of synchronized three-camera video (head and both wrists, **100** tasks × **3** views = **300** clips) and **9.7** hours of marker-based optical motion capture (**18** tasks, **288** clips) that supplies metric 3D skeletal ground truth for the same household activity classes. All streams are independently curated and annotated. The corpus is characterized along the axes most pertinent to robotics: **68%** of egocentric frames contain a visible hand,[*] the household split spans **9** activity domains and **50** goal-directed tasks while the factory split spans **10** operation families and **52** tasks, both weighted toward manipulation-dense work, and **43%** of clips are free of jump cuts and unintended scene breaks.[*] The released corpus contains **zero** duplicates across all modalities. The curation pipeline combines automated screening, vision–language model verification, and human review. Every clip ships with structured metadata and **100%** informed-consent coverage under a license that permits commercial policy training. This report documents the dataset composition across household and factory domains, quality signals, and pipeline; further implementation detail is available to independent auditors under NDA.

Keywords: egocentric video, robotics, imitation learning, world models, dataset curation, manipulation.

1 Introduction

Bumblebee is a multi-modal dataset of synchronized egocentric video, three-camera (head + dual-wrist) body-mounted video, and marker-based motion capture spanning household and factory environments, assembled for training embodied policies, world models, and manipulation systems. The guiding question throughout its construction was not how many hours we could accumulate, but

[*] Hand-visibility, lighting, and temporal-integrity (jump-cut) statistics are estimated from a random sample of 7,800 clips drawn across all household domains.

what those hours contain across what environments and from what vantage points. The dataset is built around three complementary modalities, each answering different questions in robotics training: **Egocentric video** captures the operator's first-person perspective, essential for imitation learning and egocentric policy training. **Three-camera video** adds a wrist-mounted view on each hand to the head-mounted view, giving three synchronized first-person streams per recording. This is the observation structure bimanual manipulation policies are usually conditioned on: the wrist views resolve grasp and contact detail at a scale the head camera cannot, they stay on the hand when the head view is blocked by the body or by the object being manipulated, and each hand is observed in its own moving frame of reference. The stream adds hand-level resolution rather than an external vantage point. **Marker-based motion capture** supplies metric 3D joint trajectories at millimeter precision, the reference signal that pixel-space pose estimation is trained and validated against: retargeting to humanoid and bimanual kinematic chains, physically consistent trajectory optimization, and quantitative evaluation of monocular pose predictors on the same activity classes the video streams cover. Together, the three streams form a coherent training signal: first-person behavioral intent, per-hand close-up observation, and metric skeletal ground truth.

The dataset spans two complementary domains: **Household environments** (**11,778** hours, **1,802** participants) capture everyday domestic manipulation, fine motor control, and task diversity in unstructured spaces. **Factory and industrial environments** (**2,544** hours, **1,283** clips, **482** participants) capture structured, repetitive, precision-critical operations: assembly, machining, and material handling. Together the two domains supply breadth of task diversity and depth of precision in deployment-critical settings.

The dataset is characterized along several axes relevant to robotics use: hand visibility in egocentric frames, the breadth of lighting conditions captured, coverage of varied household domains, retention of failure and recovery events, temporal integrity, multi-modal alignment, and complete provenance and consent documentation.

The corpus prioritizes properties commonly of interest for robotics: whether hands are visible, the range of lighting conditions captured, coverage of varied scenes, and the presence of rare interactions such as fumbles, corrections, and recoveries that some collection efforts discard. It also retains long recordings and documents temporal integrity across all modalities. The remainder of this report documents the dataset composition, quality signals, and curation pipeline; further implementation detail is available to independent auditors under NDA.

Contributions.

1. A multi-modal dataset of synchronized egocentric, three-camera, and marker-based motion-capture streams spanning household and factory environments: **11,778** hours of household egocentric video (**29,247** clips, **1,802** participants) and **2,544** hours of factory egocentric video (**1,283** clips, **482** participants), complemented by **75** hours of synchronized three-camera footage (**100** tasks, **300** clips) and **9.7** hours of optical motion capture (**18** tasks, **288** clips, **4** performers).
2. Egocentric hand visibility measured on a random subset of released household clips: **68%** of frames contain at least one visible hand, characterized at the video level (Section 5). Wrist-mounted views keep each hand in frame when the head view is occluded, and supply per-hand close-up observation for bimanual policy training; the motion-capture stream supplies metric joint trajectories as ground truth for pose-estimation evaluation and humanoid retargeting.
3. Environmental and lighting coverage spanning household and factory settings, with scene illumination forming a broad, well-centered distribution across all modalities rather than clustering at a single convenient exposure.
4. Temporal integrity and alignment across modalities: **43%** of clips (measured on a random subset

Figure 1. Representative frame grid across scenes, actors, and lighting conditions.

of household clips) are free of jump cuts and unintended scene breaks; the released corpus contains **zero** duplicates. Frame-accurate synchronization across egocentric, three-camera, and motion-capture streams, with timecode recorded per take.

5. Separate task taxonomies for the two egocentric domains, reported independently rather than merged: **50** household goal categories across **9** activity domains, and **52** factory tasks across **10** operation families. Household footage carries an additional tri-granular labeling of **510** verbs and **2,140** objects. Rare interactions and failure/recovery events are retained rather than filtered, and annotations span all three modalities.
6. Full consent and commercial-use licensing coverage; documented chain-of-custody from capture to release; and independent-auditor access under NDA.
7. Versioned metadata schemas and filled datasheets shipped alongside each release, covering egocentric video, three-camera video, and motion-capture modalities.

The remainder of this report is structured as follows. Section 2 gives a complete quantitative portrait of Bumblebee across the three major axes. Sections 3–5 document lighting, composition, and hand visibility in depth. Section 6 describes the collection protocol. Section 7, the curation pipeline. Section 8 documents the VLM and human-in-the-loop annotation pipeline and the annotation layer it releases. Sections 9 and 10 cover metadata, ethics, consent, and licensing. Section 11 sets out the immediate next release, and Section 12 the longer-horizon roadmap for capture modalities, long-tail coverage, and collection control. Appendices carry the full task distribution data for household (Appendix A), factory (Appendix B), three-camera (Appendix C), and motion capture (Appendix D).

Figure 2. Three-camera capture. Each row is a different bimanual manipulation task; the **3** frames within a row are the same time step seen simultaneously from all **3** synchronized views, the head-mounted camera and one camera on each wrist.

2 The Corpus at a Glance

Beyond hours, this section characterizes what the corpus contains across three synchronized modalities: visible hands at work, environments and lighting spanning a range of conditions, temporal segments long enough to carry task structure, and clips that are distinct from one another.

Multi-modal composition. Bumblebee comprises three synchronized data streams across household and factory environments. The **egocentric stream** contains **11,778** curated hours across **29,247** household clips and **2,544** hours across **1,283** factory clips, recorded in first-person perspective. The **three-camera stream** provides **75** curated hours across **300** clips, recorded from three synchronized body-mounted cameras, one head-mounted and one on each wrist, giving per-hand close-up observation alongside the head view (Figure 2). Throughout this report, "Tri-Camera Stream" (equivalently, the *head + dual-wrist stream*) refers to this configuration: one head-mounted camera plus one wrist-mounted camera per hand. Each of the **100** tasks is captured from all **3** views, so **100** tasks $\times$ **3** views = **300** clips. The **motion-capture stream** provides **9.7** hours (**580** minutes) across **288** approved clips of marker-based optical capture, delivering metric 3D joint trajectories over **18** household manipulation and locomotion tasks (Figure 3). Every task is performed by each of the **4** performers, so **18** tasks $\times$ **4** performers = **72** takes, and each take is captured from **4** synchronized views, so **72** takes $\times$ **4** views = **288** clips. All streams are temporally aligned and independently curated.

Scale and capture (Egocentric). Capture spans the full consumer-to-prosumer fidelity range: native resolutions from **FHD–4K** and frame rates of **30–120 fps**, so downstream teams can train and evaluate across the resolutions and shutter rates they actually deploy on. Every clip carries structured metadata versioned alongside the data release, with synchronization metadata linking egocentric, three-camera, and motion-capture recordings at frame-level precision.

Figure 3. The three released motion-capture streams, one per row: 3D pose render (top), solved skeleton (middle), and RGB reference video (bottom), with five samples drawn from each.

Hand visibility. Hand visibility is a common filter for egocentric footage, so we report it directly. **68%** of frames (measured on a random subset of household clips) contain at least one visible hand; Section 5 covers this in more detail. The task vocabulary spans **510** verbs and **2,140** objects across **50** goal categories, with rare interactions preserved rather than filtered.

Environmental coverage (household). The household split is captured in real domestic environments and spans **9** activity domains: cleaning, organization, cooking and food preparation, laundry management, daily routines, assistance and navigation, preparation and occasions, hosting, and home maintenance. Footage covers common household spaces such as kitchens, bathrooms, bedrooms, living rooms, and storage and utility areas. Every domain is manipulation-dense by construction: the corpus targets hands-on household work rather than passive observation, and the per-domain breakdown is given in Table 2.

Operational coverage (factory). The factory split spans **52** distinct tasks recorded on active production lines, grouped into **10** operation families by the manipulation skill each station demands rather than by location. Table 3 gives the per-family breakdown.

Lighting and optical conditions. We measure average scene illumination per clip as mean frame luminance. Across the corpus the distribution is bell-shaped and centered near the middle of the normalized luminance range (Figure 4), with meaningful mass in both the darker and brighter tails, indicating coverage across a range of lighting conditions rather than concentration at a single exposure level.

Temporal structure. Mean clip duration is **24** minutes, with **36%** of hours in long-horizon episodes of at least **40** minutes. **43%** of clips (measured on a random subset of household clips) are free of jump cuts and unintended scene breaks; the remainder ship with exact cut timestamps. **Every cut identified in QC is documented with an exact timestamp rather than left implicit.**

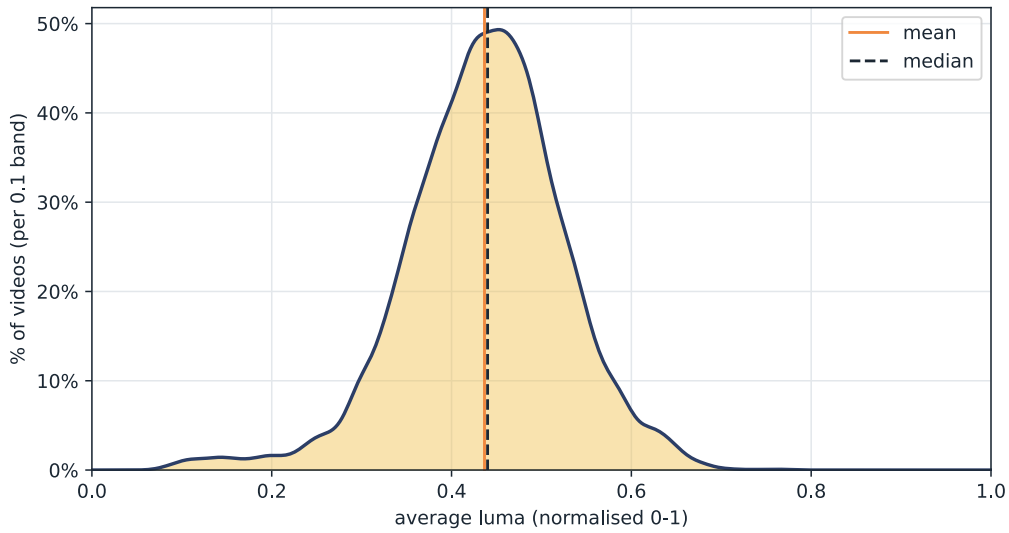


Figure 4. Scene-brightness distribution across the corpus: share of videos by average luma (normalized 0–1), with reference lines at the corpus mean and median.

Uniqueness and integrity. Duplicate detection was run across the corpus. The released corpus contains **0** duplicates. **218** duplicate candidates were removed before release.

Provenance and consent. **100%** of clips ship with informed consent covering research and commercial use. Bystanders and personally identifying content are handled per jurisdiction-appropriate policy; released frames contain no identifiable non-consenting individuals. Chain-of-custody is documented from capture to release.

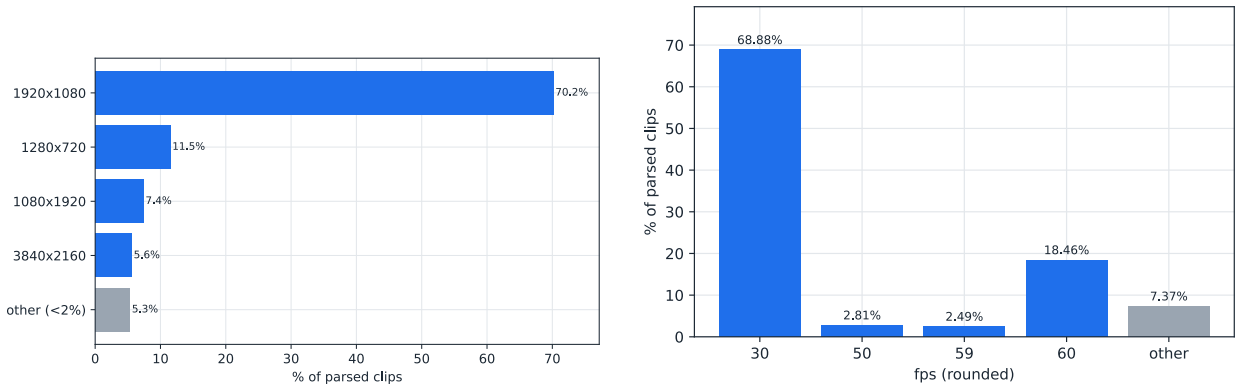
3 Lighting and Optical Conditions

Lighting varies widely in deployment settings. Rather than binning clips into coarse brightness categories, we measure scene brightness directly as mean frame luminance, computed per frame and summarized per video. Figure 4 shows the resulting distribution across the corpus. It is bell-shaped and centered roughly at the middle of the normalized luminance range, with substantial mass extending into both the darker and brighter tails. In other words, the everyday span of indoor and outdoor lighting is represented across the corpus, and the data is not concentrated at a single exposure level.

4 Composition and Diversity

4.1 Household scene distribution

The household split spans **9** activity domains, weighted toward the everyday cleaning, organization, and cooking work that dominates domestic settings. Table 2 gives the full per-domain breakdown by recorded hours and clips. The distribution is intentionally non-uniform, following where hands-on manipulation is densest rather than sampling domains uniformly.



(a) Native resolution.

(b) Frame rate.

Figure 5. Capture format across the parsed corpus: native resolution (share of clips) and frame rate (share of clips).

4.2 Factory operation-family distribution

Table 3 gives the per-family breakdown of the **52** factory tasks; the full per-task listing is in Appendix B. Three families, component insertion and assembly, inspection and quality verification, and precision soldering and joining, together account for **65%** of factory hours. That concentration is deliberate: these are the tight-tolerance, high-repetition operations where manipulation-policy performance matters most and where human demonstrations are hard to substitute.

4.3 Capture format

Clips are released at their native capture resolution and frame rate rather than transcoded to a common target, so downstream teams can select the exact fidelity their pipeline expects and study robustness to capture conditions instead of having it flattened away. Resolutions span the full-HD-to-4K range, and frame rates cover 30, 50, 60, and 120 fps, from standard-motion footage through high-frame-rate capture that resolves fast hand motion without blur. Figure 5 shows both distributions across the parsed corpus.

4.4 Task and interaction taxonomy

Household tasks are labeled at three granularities: verb (*fold*), object (*shirt*), and goal (*fold and put away laundry*). The household split comprises **50** distinct tasks grouped into the **9** activity domains of Table 2.

The factory split comprises **52** distinct tasks grouped into the **10** operation families of Table 3. Factory tasks carry station-level identity rather than the verb/object/goal decomposition applied to household footage, because a factory task name already denotes a fixed, repeatable operation on a known part. Across both splits, failure and recovery events, including dropped objects, mis-grasps, and corrections, are retained rather than filtered, a category that curated imitation-learning data often removes.

4.5 Temporal structure

Mean clip duration is **24** minutes. The distribution is bimodal: short manipulation vignettes (**88%** of clips, mean **18** min) and long-horizon episodes (**12%** of clips, mean **71** min). The latter carry **36%** of total hours. Figure 6 shows the full clip-duration histogram.

Temporal integrity is measured directly rather than assumed: clips are scored for jump cuts and

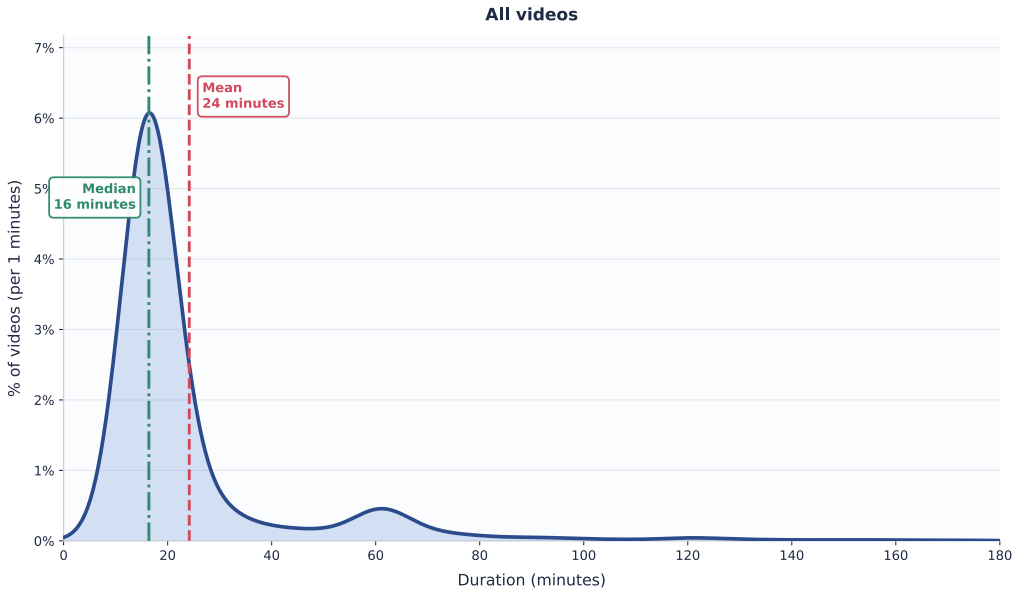


Figure 6. Clip-duration distribution.

unintended scene breaks. **43%** of clips (measured on a random subset of household clips) contain no jump cuts, and clips with cuts carry them by design and are documented with exact timestamps.

4.6 Long-horizon episodes

Long-horizon episodes are recordings of **40** minutes or more. A single recording of this length carries a full multi-step task, setup through completion, at a scale relevant to world-model and long-context imitation training. We count every video of at least **40** minutes as long-horizon.

Bumblebee contains **3,593** long-horizon videos, **12%** of all clips, together carrying **4,225** recorded hours, or **36%** of the corpus's total. Figure 7 shows the duration distribution across all long-horizon videos: recording lengths extend well beyond the **40**-minute cut-off into multi-hour material, giving downstream teams a substantial long-context base rather than a handful of outliers.

5 Hand Visibility

Hand visibility is one of the most common filters teams apply to egocentric footage, so we measure it rather than assert it. Within a random subset of household clips every frame is scored for hand presence, and those frame scores are aggregated per video. **68%** of frames in that subset contain at least one visible hand.

Both-hands visibility is the more demanding measure, and is often what teams working on two-handed tasks want to filter for. In the same subset, **45%** of videos have both hands visible in more than **20%** of their frames. We report this as a visibility statistic only: it reflects how often both hands appear in frame, not a judgment about the manipulation being performed.

The synchronized three-camera stream addresses occlusion differently: because a camera is mounted on each wrist, the hand remains in frame even when the head view is blocked, so hand visibility does not depend on head orientation (Figure 2). The motion-capture stream closes the loop on this: for the **18** task categories it covers, marker-based joint trajectories provide metric ground truth against which pose estimates recovered from either video stream can be scored directly, rather than evaluated against pseudo-labels produced by another vision model. Together the three streams

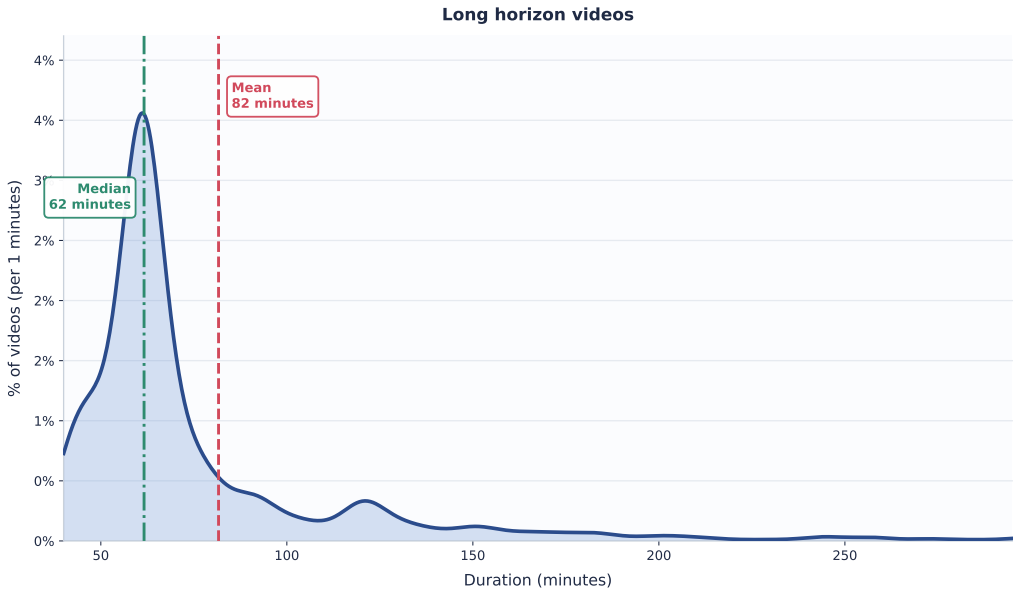


Figure 7. Duration distribution across all long-horizon videos (at least **40** minutes). Durations are shown in minutes.

ground learning in first-person behavioral intent, per-hand close-up observation, and metric skeletal structure. Alignment across all streams is frame-accurate and documented in the synchronized metadata.

6 Collection Protocol

We describe collection at the level required to interpret the corpus, not at the level of operational reproducibility. Recording devices, participant sourcing partners, task-prompt structure, and per-participant compensation are not disclosed here.

Hardware and modalities. Three synchronized data streams were captured:

- **Egocentric video:** Captured at native resolutions from **FHD–4K** at frame rates of **30–120 fps**.
- **Three-camera video:** Three body-mounted cameras per recording, one head-mounted and one on each wrist, running synchronized. All three views are first-person; the wrist views observe each hand and the manipulated object at close range.
- **Marker-based motion capture:** Optical marker tracking in a calibrated volume, producing metric 3D joint trajectories per performer. Sessions were recorded with synchronized reference video and per-take timecode. Multi-performer takes were captured for interaction tasks (*e.g.,* coordinated object transfer), so the stream covers both single-agent and two-agent manipulation.

All streams are temporally synchronized to frame accuracy and released alongside one another. Calibration and synchronization parameters are included in the clip metadata.

Participant recruitment and consent. Participants were recruited through Deccan AI Experts against inclusion criteria that prioritized geographic diversity and coverage of manipulation-dense professional contexts. Motion-capture sessions were performed by a dedicated cohort of **4** trained performers, each executing every task category so that per-task variation across performers is measurable rather

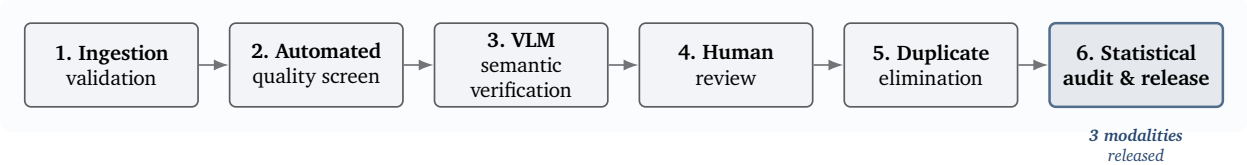


Figure 8. Curation pipeline. Each stage is applied in sequence to every candidate clip; per-stage attrition is documented and released to licensees.

than confounded with task identity. Every participant signed informed consent covering research and commercial use of the resulting data across all three modalities before any recording took place.

Task prompting. Participants received naturalistic task prompts calibrated to the scene category. Prompts were designed to elicit realistic, self-directed activity rather than staged performance; specific prompt structure is not disclosed. Failure events were explicitly retained rather than requested to be re-recorded.

Bystander policy. All recording sessions were conducted in accordance with jurisdiction-appropriate bystander policies across all capture modalities. Motion-capture sessions were conducted in a closed calibrated volume with no bystander presence.

7 Curation Pipeline

We describe the curation pipeline in enough detail to interpret the reported numbers, without publishing its implementation. The pipeline processes all three synchronized modalities: egocentric, three-camera, and motion capture, maintaining consistency and integrity across streams. Specific model choices, prompt structures, thresholds, and internal review rubrics are not disclosed; independent auditors are granted access under NDA. This balance follows common practice for commercially licensed datasets: enough detail to interpret and trust the numbers, without releasing an implementation that could be replicated wholesale.

7.1 The funnel

The pipeline is a funnel. Every clip across all modalities that ships passed the six stages in Figure 8 in order, with attrition documented per stage, so that the material screened out at each step is accounted for rather than hidden. Modality-specific screening supplements the shared stages: per-view synchronization and framing checks for the three-camera stream, and marker-occlusion and solve-quality review for motion capture, where every recording carries an explicit approval status and only approved recordings are released. Clips are retained only if they pass curation in every modality in which they appear.

7.2 Automated quality screening

Stage 2 rejects clips on format-independent quality signals. Rejection triggers include excessive motion blur, sustained under- or over-exposure, incoherent ego-motion trajectories, audio-track corruption where audio is expected, and encoding artifacts. Per-clip quality scores are attached to every clip that survives this stage and are released as metadata.

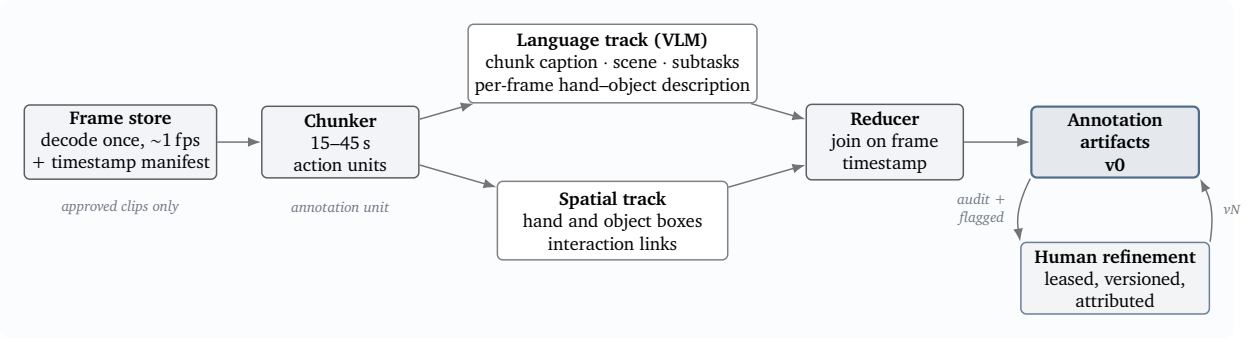


Figure 9. Annotation pipeline. Language and spatial tracks run concurrently over the same frame store and are joined on frame timestamp; human refinement writes new versions on top of the machine output rather than replacing it.

7.3 VLM-based semantic verification

Stage 3 applies a vision–language model to every surviving clip. Its outputs are used both to filter (*e.g.,* verifying that declared scene labels match observed content, that hand presence claims are consistent with the visual signal, and that safety-relevant properties hold) and to enrich metadata with structured semantic annotations. The model is calibrated against human-in-the-loop review: its decisions are audited by reviewers, and low-confidence cases are escalated to them (Section 7.4). The same infrastructure produces the shipped semantic annotation layer; Section 8 describes it. We do not disclose the specific model, prompt structures, or decision thresholds.

7.4 Human review

Stage 4 applies human review in two modes: uniform random-sample audit for pipeline calibration, and targeted review of any clip flagged by upstream stages. Reviewers work against a rubric that is versioned alongside the metadata schema but is not published.

7.5 Duplicate elimination

Stage 5 removes duplicate clips, detecting both re-encodings and format conversions and clips that share substantial content despite differing in framing, cropping, or compression. **218** candidates were removed at this stage. The released corpus contains **0** duplicates.

8 VLM and Human-in-the-Loop Annotation

Curation decides which clips ship. Annotation decides what each shipped clip *says*: what the wearer is doing, at what granularity, with which hand, on which object. That layer is produced by a dedicated annotation pipeline (Figure 9) that runs only on clips already approved by curation, one job per episode, event-driven, so annotation compute never sits on the critical path of collection or contributor payment.

The pipeline pairs a vision–language model with human refinement. The model provides breadth: every clip, every frame, at a scale that would be impractical to cover by review alone. Trained annotators set the standard of correctness: they audit, correct, and version the machine output, and their judgments are what the model's prompts and thresholds are calibrated against. Neither half is treated as ground truth on its own.

8.1 Frame preparation and chunking

Each approved episode is decoded once into a frame store at roughly one frame per second, accompanied by a manifest carrying the exact timestamp of every stored frame. Every annotation downstream is keyed to a frame index and its manifest timestamp, never to a container frame count: source containers are routinely fragmented and report frame counts that are simply wrong, and an annotation layer that trusts them silently drifts against the video.
The timeline is then split into contiguous action chunks of 15–45 seconds (30 s target), merging a short trailing window into its predecessor. The chunk is the unit of temporal reasoning: it is long enough to contain a complete action with its approach and release, and short enough that a single description of it stays specific.

8.2 Two-scale vision–language reasoning

The language track runs two call patterns concurrently over the same frames, at two different granularities.

Chunk scale. A temporally ordered subsample of each chunk's frames, with their timestamps, is sent in one call. The model returns a caption of what happens across the segment, a scene classification drawn from a closed vocabulary with a confidence, and a segmentation of the chunk into one to six contiguous, non-overlapping subtask intervals (*reach*, *grasp*, *move*, *place*, *release*, *adjust*), each with its own start, end, label, and confidence. Idle stretches are labeled as such rather than being forced into an action label.

Frame scale. Frames are tiled into batches and annotated with one short natural-language description each: what the hands are doing and which objects or surfaces they touch or hold, or an explicit statement that no hands are visible. This is the layer that makes the corpus searchable at the level of contact events rather than clips.

Egocentric-specific constraints. First-person footage has a failure mode that generic captioning prompts walk straight into: hand-side inversion. Handedness is resolved from where the hand sits in the frame, with no dominant-hand default and an explicit unlabeled fallback ("a hand") when the side is genuinely ambiguous. Getting this wrong is invisible in aggregate statistics and poisonous to any policy that learns bimanual roles.

Structured output and degradation. Responses are requested as strict JSON and parsed defensively: subtask spans are clamped to their chunk, confidences are clamped to a unit interval, malformed items and unknown frame indices are dropped rather than failing the call, and complete entries are salvaged from truncated replies. Over an episode with hundreds of batched calls, at least one imperfect reply is close to certain; the pipeline is built so that this costs a few frame descriptions rather than an episode.

8.3 Spatial track

Boxes come from a dedicated egocentric hand–object interaction detector running on GPU, not from the vision–language model. That is a measured decision rather than an architectural preference: the language model's spatial grounding was not at shippable quality, while a specialist detector produces per-frame boxes for both hands and the first and second manipulated objects, plus explicit hand → object interaction links with probabilities, at no per-call inference cost. Boxes are normalized to frame coordinates, track identifiers carry the role (left hand, right hand, first object), and the open-vocabulary object names come from the language track, so naming is paid for once per chunk

rather than once per box. An optional mask-propagation pass tightens boxes to object contours where pixel-accurate extents are needed.
The tracks are independent and fail soft. A spatial-track failure is recorded and the episode still ships its language layer; each episode's annotation manifest records which tracks actually ran, with model versions and timings, so no consumer has to infer coverage from the presence or absence of fields.

8.4 Reduction and artifacts

The reducer joins per-frame and per-chunk outputs on frame timestamp and writes three artifacts per episode: a frame table with one row per decoded frame (detections plus the hand–object description), one record per chunk (caption, scene, subtasks), and a manifest recording tracks, model versions, timings, and counts. One row is emitted for every manifest frame even when both tracks are empty, so consumers never have to special-case gaps. The manifest is written last, which makes its presence the completion marker and reruns idempotent.

8.5 Human-in-the-loop refinement

Machine output is version 0 and is never overwritten. Annotators work in a player that renders the annotations directly over the video: box overlays with interpolation between stored frames, a clickable subtask timeline, and inline caption, scene, and hand-description editors. Editing is keyboard-first, so that frame-level corrections stay quick and comfortable to make. The supported operations are box move, resize, add, delete, and propagate-forward; label rename with autocomplete over labels already seen in the episode; subtask boundary adjustment, relabeling, splitting, and merging; caption, scene, and hand-description rewriting; and marking an episode reviewed.
Consistency when several annotators work at once is handled by the system rather than left to coordination between them. Each episode carries a short-lived single-writer lease; every save is a single transaction conditioned on both a live lease and the base version the annotator actually loaded; edit identifiers are generated client-side so a retried save is applied once rather than twice; and a save based on a stale version is surfaced as a conflict, so the annotator can reapply their edit on the current version instead of having it silently merged. Version bodies are append-only and attributed, so any released annotation can be traced back through every human edit to the machine version it started from, with credit to the people who made it.
Annotator attention is directed the same way it is in curation: a uniform audit sample to keep the pipeline calibrated, plus targeted review of low-confidence and flagged episodes. Disagreements found in review are what drive revisions to prompts and thresholds, which is why annotation review is treated as part of the pipeline rather than a quality check appended to its end.
As with curation, specific models, prompt text, vocabularies, and thresholds are not published; independent auditors are granted access under NDA.

8.6 The released annotation layer

The layer ships against the same versioned schema as the rest of the metadata, additively: per-chunk captions, scene classifications, and subtask segmentations with confidences; per-frame hand–object descriptions; and, where the spatial track ran, per-frame hand and object boxes with hand → object interaction links. Each episode carries a manifest naming which annotation tracks ran and the model versions that produced them, and refined episodes carry the annotator attribution and version history behind them.
Machine and human annotations are shipped side by side rather than collapsed into a single track, because the two carry different information. Figure 10 shows the review tool on three household episodes: hand boxes overlaid on the frame, the model's chunk-level caption, scene class, and

Figure 10. The annotation layer as reviewers see it, on household episode: repositioning a bucket between rooms. The panel shows per-frame left and right hand boxes over the video, the model's chunk caption with its scene class and timestamped subtask breakdown, and the human-written chunk for the same span where one exists. The two tracks at the foot of the panel are the human and model chunk timelines for the full episode; their differing density is the granularity enhancement the human pass provides.

timestamped subtask breakdown, and above it the human chunk covering the same span, written at a finer granularity that names the grasp itself (which fingers oppose which, and on which part of the object). The two timelines at the bottom of each panel make the difference in granularity explicit: the model segments the episode into fixed windows, while the human track follows the actual action boundaries and is denser where the manipulation is. Where a human chunk has not been written yet, the panel says so rather than silently substituting the model's version.

9 Metadata

Every clip in all three modalities ships with structured metadata versioned alongside the data release. Egocentric metadata includes timestamps, scene labels, lighting tags, hand-visibility flags, task labels at three granularities, and per-clip quality scores. Three-camera metadata adds the mount position of each view (head, left wrist, right wrist), per-view hand and object visibility flags, and the inter-view synchronization offsets. Motion-capture metadata adds per-take timecode in and out points, task and performer identity, take and approval status, subject count for multi-performer interaction takes, and the synchronized reference-video reference. Synchronization metadata links all streams at frame-level precision, enabling cross-modal queries and multi-stream training pipelines. The semantic annotation layer of Section 8.6 is carried alongside these fields under the same schema version.

10 Ethics, Consent, and Licensing

All participants signed informed consent covering research and commercial use of the resulting data across all three modalities (egocentric video, three-camera video, and marker-based motion capture) prior to any recording. No frames containing minors are included unless explicit guardian consent was recorded and jurisdictional requirements met. Bystander and personally-identifying

content are handled per jurisdiction-appropriate policy; the released corpus contains no identifiable non-consenting individuals in any released frame across any modality.
Full licensing terms are available separately. The default license supports commercial model training and downstream model release under standard commercial terms. Independent audits of consent coverage, redaction, and provenance across all three modalities are available to licensees under NDA.

11 Next Release

This release characterizes the corpus at the level of clips, scenes, and tasks. The next phase moves one level deeper, to the interactions themselves.

Hand–object interaction analysis. The immediate next deliverable is a dedicated hand–object interaction report covering the released corpus: which objects are contacted, when contact begins and ends, how each grasp is formed and released, and how bimanual coordination is distributed across tasks. Contact structure is the layer that separates footage a policy can learn manipulation from and footage that merely shows a task being completed, and it is the axis teams most often have to annotate themselves before egocentric data becomes trainable. We are measuring it directly on Bumblebee rather than leaving it to downstream inference, and we will report it with the same discipline applied here: measured on the released data, with the sampling basis of every statistic stated alongside it.
Because the analysis runs over the corpus already released, it requires no re-collection and no change to the data licensees hold. It ships as an additive layer against the same versioned metadata schema.

12 Future Work and Improvements

Bumblebee is a living corpus: collection continues beyond this release. Four directions are already in motion.

Richer capture modalities. The current release spans egocentric video, the Tri-Camera Stream, IMU, and marker-based motion capture. We are extending this set with the signals robot-manipulation teams ask for most. Instrumented tactile gloves record what the hands feel: contact, grip force, and finger pose. UMI-style handheld grippers record demonstrations from the gripper's point of view rather than the wearer's, which makes them easier to transfer to real robot arms. Eye-gaze tracking shows where the wearer is looking, a strong signal of attention and intent. Denser depth and higher-rate IMU streams improve 3D reconstruction. Every new modality plugs into the same synchronization and metadata schema as the current streams, so the corpus grows as one dataset rather than splitting into incompatible pieces.

Long-tail environments. Raw hours are no longer the bottleneck; coverage of rare situations is. We are using the corpus's own metadata and clip embeddings to find environments that appear less often in our data than they do in the real world: cramped or dimly lit kitchens, shared and multi-generational households, festival preparations, aging and non-standard appliances, and unusual factory cells. Each gap becomes a concrete collection brief, so rare settings are captured deliberately instead of by luck.

Synthetic data, teleoperation, and the data pyramid. We think of embodied training data as a pyramid. The base is large-scale human data: web video and text alongside egocentric recordings of people doing real tasks. Bumblebee sits at the high-quality end of this base layer. Unlike scraped web video, it is first-person, synchronized, measured, and task-directed, which is what makes the

base usable for manipulation training. The middle layer is synthetic data: simulation and generative augmentation that multiply coverage cheaply, but only approximate real physics and appearance. The top is data from real robots: teleoperated demonstrations and autonomous rollouts, which match the target hardware exactly but are the most expensive to scale. Our plan follows the pyramid. We keep growing the structured human base, add simulation and synthetic variants for settings that are impractical to capture physically, and extend toward teleoperation on common robot platforms using the same task list. Because every layer shares one task taxonomy and metadata schema, data from one layer can train and validate models built on another.

Adaptive distribution control. Today we keep the corpus balanced through collection briefs and after-the-fact curation. We are moving to a closed loop. Our dashboards already show, in near real time, how much each task and scene contributes to the overall corpus. We are connecting those numbers back to task assignment: once a category reaches its target share, contributors are automatically redirected to categories that are still short. The corpus then stays balanced while it is being collected, and drift shows up as an alert during collection instead of a surprise at release.

Acknowledgements

Bumblebee is the product of many hands beyond the author list, and we are deeply grateful to the people whose day-to-day work made it real.

We thank our Operations team, namely **Vishwas Choudhary, Glady Franklin.S, Aditya Rai, Ketan Agrawal, B K Lakshita Yadav, Kannegulla Sai Vardhan, and Karthikeya Siripuram**, for their valuable contributions in running collection at scale: sourcing and managing collection partnerships, recruiting and briefing participants across household and factory sites, and keeping recordings flowing end to end, week after week.

We are equally grateful to our Platform team, namely **Rituraj Sarkar, Soubhik Biswas, and Sandeep YSV**, for designing and building the platform that captured and ingested every recording, and for keeping our software and ML systems healthy throughout, from capture tooling to the curation pipeline with which every number in this report was produced.

We also thank our Dashboard team, namely **Pavan Bhamidpati and Aishwarya Karampuri**, together with everyone who builds and operates the platform, for instrumenting the full Bumblebee corpus end to end: every recording, its metadata, and its quality signals are indexed and served through interactive dashboards that give our technical and go-to-market teams a single pane of glass for exploration, review, and audit.

Finally, we thank the participants who welcomed our cameras into their homes and workplaces. The scale, diversity, and integrity of this release would not have been possible without all of them.

Appendix A Household task distribution

Per-task composition of the household egocentric data, grouped by activity domain and ordered by recorded hours within each domain. Clips are approved recordings; hours are approved recorded hours.

Task	Clips	Hours
Cleaning		
Deep cleaning & polishing a collection of shoes	407	330

continued on next page

Task	Clips	Hours
Stovetop & burner deep clean	937	272
Refrigerator shelf & compartment deep clean	835	270
Deep cleaning behind & beside major appliances	331	267
Disinfecting doors, handles & light switches	844	254
Living-room floor mopping	893	251
Kitchen countertop & appliance cleaning	846	247
Outdoor / balcony furniture cleaning	287	232
Windows & window sills	771	227
Living-room surface clean	789	222
Bathroom deep clean: sink & toilet	793	220
Bathroom mirror, countertop & floor clean	765	214
Pantry / dry-storage shelf deep clean	692	207
Curtains: take down, wash & re-hang	242	203
Ceiling-fan blades	660	185
Under & behind furniture	634	181
Under-sink cabinet clean & organize	609	169
Window-AC unit deep clean	218	169
Trash bins & liner replacement	468	128
Subtotal	**12,021**	**4,248**
Organization		
Changing bed sheets & making the bed	1,052	294
Study / work desk setup	872	251
Household documents, bills & papers	308	236
Kitchen utensil sorting & cabinet organization	836	233
Charging & cable-management station	349	220
Bedroom closet & shoe organization	765	211
Medicine cabinet & first-aid storage	723	211
Kids' study / activity corner	310	211
Grocery stocking	631	181
Living-room setup & organization	652	180
Seasonal storage sort	199	159
Subtotal	**6,697**	**2,387**
Cooking & food preparation		
Curry prepared from raw ingredients	411	399
One-pot rice dish	393	357
Homemade snack or starter	381	337
Traditional dessert	362	283
Batch-cooking ingredient prep & kitchen setup	261	255
Essential kitchen food prep	611	189

continued on next page

Task	Clips	Hours
Subtotal	**2,419**	**1,820**
Laundry management		
Ironing clothes	1,008	369
Folding clothes & putting away	1,080	321
General laundry handling	884	244
Laundry prep: sort, carry & load washer	861	238
Bathroom linen turnaround & toiletry organization	655	183
Subtotal	**4,488**	**1,355**
Daily routines		
Morning household routine, bed to breakfast	385	333
Evening wind-down, kitchen to bedroom reset	324	279
Subtotal	**709**	**612**
Assistance & navigation		
Multi-room navigation & item collection	720	196
Indoor plant care	615	175
Kitchen assistance: pour, open, microwave	598	167
Subtotal	**1,933**	**538**
Preparation & occasions		
Packing for a day outing or picnic	270	226
Festival / occasion home preparation	175	155
Subtotal	**445**	**381**
Hosting		
Setting up the house for guests	305	258
Subtotal	**305**	**258**
Home maintenance		
Home-maintenance round: bulbs, fixtures & batteries	230	179
Subtotal	**230**	**179**
Total	**29,247**	**11,778**

Appendix B Factory task distribution

Per-task composition of the factory egocentric data, grouped into the **10** operation families of Table 3 and ordered by recorded hours within each family. The **52** tasks are distinct manufacturing and assembly operations captured on active production lines. The split comprises **1,283** clips in total, but clip counts were not recorded per task, so hours are reported here without a clip column.

Task	Hours
Component insertion and assembly	
Manual insertion	319
Manual assembly	182
Pick and place	30
Remote assembly	27
Cable preparation	2
Subtotal	**560**
Inspection and quality verification	
Visual inspection	208
Final quality check and visual inspection	206
Post-wave soldering inspection	41
Pre-wave soldering inspection	32
Pre-reflow inspection	23
Automated optical inspection (AOI)	15
Reflow inspection	13
Raw-material sampling and quality check	9
Receiving and final PCBA inspection	2
Subtotal	**549**
Precision soldering and joining	
Manual soldering and touch-up	478
Gluing and soldering	39
Touch-up	19
Potting	10
Subtotal	**546**
Electrical test and debug	
Functional circuit test	146
Testing	49
Shorting	33
Remote testing	20
RAM debug	19
PCBA programming	10
Debug rework	6
Burn-in area debugging	5
Subtotal	**288**
Packaging and kitting	
Packing	95
Kitting	34
Box making	18

continued on next page

Task	Hours
Battery sealing and packing	8
Taping	5
Subtotal	**160**
Material processing and forming	
Cutting	59
Preforming	31
Bending	24
PCB depaneling	23
MOSFET bead forming	13
Laser engraving	5
Subtotal	**155**
Machine tending and material handling	
Machine-assisted sorting and screwing	48
Loading and unloading	30
PCB loading and scanning	19
Scanning	17
Weighing	6
Subtotal	**120**
Surface preparation and cleaning	
PCB cleaning	24
Stacking and cleaning	18
Controller masking	15
Remote cleaning	8
Subtotal	**65**
Marking and labeling	
Marking on LCD	33
Labeling	23
PCB printing	8
Subtotal	**64**
Line supervision and unclassified	
Humming	17
Unclassified station work	14
Line supervision	6
Subtotal	**37**
Total	**2,544**

Appendix C Three-camera task distribution

Per-task composition of the three-camera data. Tasks are standardized bimanual manipulation operations, each capturing multiple demonstrations of a single dexterous skill under a fixed protocol. Every task is captured from all **3** views, so the Clips column is **3** per task and **100** tasks × **3** views = **300** clips. The Hours column is the footage summed across the **3** views.

Task	Clips	Hours
1. Slide-open ziplock and insert item	3	0.75
2. Seal ziplock along full length	3	0.75
3. Zip-tie cable/pen bundling	3	0.75
4. Trim zip-tie tail with scissors	3	0.75
5. Thread zip tie through two holes	3	0.75
6. Tear adhesive tape and apply to paper edge	3	0.75
7. Wrap adhesive tape around a small box	3	0.75
8. Load a new tape roll into dispenser	3	0.75
9. Painters-tape straight-edge masking	3	0.75
10. Peel painters tape cleanly	3	0.75
11. Cap and uncap a pen	3	0.75
12. Click and align retractable pens	3	0.75
13. Cut paper along a drawn line	3	0.75
14. Cut uniform paper strips freehand	3	0.75
15. Score and snap paper along a ruler	3	0.75
16. Precise half-fold and crease	3	0.75
17. Accordion / airplane fold sequence	3	0.75
18. Paper-clip a document stack	3	0.75
19. Link paper clips into a chain	3	0.75
20. Binder-clip onto a paper stack	3	0.75
21. Remove and refit binder-clip arms	3	0.75
22. Push-clip fastener into punched holes	3	0.75
23. Rubber-band bundle of pens/sticks	3	0.75
24. Stretch rubber band over a container	3	0.75
25. Twist-tie a bag neck closed	3	0.75
26. Untwist and re-twist standalone ties	3	0.75
27. Measure an object and lock the tape	3	0.75
28. Controlled tape-measure retraction	3	0.75
29. Extend tape and mark length on paper	3	0.75
30. Peel and place sticky notes aligned	3	0.75
31. Stack and square loose sticky notes	3	0.75
32. Insert paper into envelope and seal	3	0.75
33. Open a sealed envelope with a blade	3	0.75
34. Lace string through a card	3	0.75

continued on next page

Task	Clips	Hours
35. Tie a knot / bow with string	3	0.75
36. Wind string onto a card/spool	3	0.75
37. Thread beads onto a string	3	0.75
38. Thread a nut onto a bolt by hand	3	0.75
39. Screw and unscrew a small jar lid	3	0.75
40. Stack coins and drop into a slot	3	0.75
41. Roll coins into a paper sleeve	3	0.75
42. Clip clothespins onto an edge	3	0.75
43. Assemble and disassemble small blocks	3	0.75
44. Cut bubble wrap and wrap an item	3	0.75
45. Hook-and-loop strap a cable bundle	3	0.75
46. Transfer small items with tweezers	3	0.75
47. Draw and dispense water with a syringe	3	0.75
48. Wind a rubber-band ball	3	0.75
49. Fold-assemble a small cardboard box and tape	3	0.75
50. Peel adhesive label and apply to a bottle	3	0.75
51. Punch and fasten a booklet	3	0.75
52. String-bound mini-notebook	3	0.75
53. Gift-wrap a small box	3	0.75
54. Band-and-flag a pen bundle	3	0.75
55. Envelope stuffing and clip	3	0.75
56. Coin sorting into labeled cups	3	0.75
57. Nut-on-bolt into jar	3	0.75
58. Bead loop on pipe cleaner	3	0.75
59. Measure, cut and curl ribbon	3	0.75
60. Measure and cut string to length	3	0.75
61. Zip-tie mount to a card	3	0.75
62. Roll and band paper tubes	3	0.75
63. Load battery and secure lid	3	0.75
64. Syringe transfer between jars	3	0.75
65. Transfer cotton balls into a jar with tweezers	3	0.75
66. Peg documents along a line	3	0.75
67. Origami cup fold	3	0.75
68. Jog and band a paper ream	3	0.75
69. Cross-band a stack	3	0.75
70. Bubble-wrap and label a bottle	3	0.75
71. Grid zip-tie bundle	3	0.75
72. Label-grid application	3	0.75
73. Twist-tie and clip bags	3	0.75
74. Pack and tape a box	3	0.75

continued on next page

Task	Clips	Hours
75. Paper chain construction	3	0.75
76. Zig-zag lace and tie-off	3	0.75
77. Sticky-note flag indexing	3	0.75
78. Pen disassembly and reassembly	3	0.75
79. Note-under-band tuck	3	0.75
80. Double-sided tape mounting	3	0.75
81. Measured drop dispensing	3	0.75
82. Coil and strap a cable	3	0.75
83. Binder clips onto a rod row	3	0.75
84. Free-standing card structure	3	0.75
85. Twist-tie pattern through card	3	0.75
86. Kit assembly with bag and tag	3	0.75
87. Round-corner card cutting	3	0.75
88. Coin stack-sort then bank	3	0.75
89. Insert brad fasteners in a grid	3	0.75
90. Tie a luggage-style tag	3	0.75
91. Sleeve insert and clip	3	0.75
92. Paper cone / funnel	3	0.75
93. Sort mixed hardware	3	0.75
94. Bundle pens with string and knot	3	0.75
95. Peel and stage tape tabs	3	0.75
96. Fold, insert and label a mailer	3	0.75
97. Wind ribbon into a bow and clip	3	0.75
98. Bead color-pattern threading	3	0.75
99. Stack and clip index cards	3	0.75
100. Nested cup stacking	3	0.75
Total	**300**	**75**

Appendix D Motion-capture task distribution

Per-task composition of the marker-based motion-capture stream, ordered by recorded duration. Every task category was performed by all **4** performers, so **18** tasks × **4** performers = **72** takes; each take was captured from **4** synchronized views, so **72** × **4** = **288** approved clips carrying **580** minutes of metric 3D joint trajectories. The Clips column is therefore **4** performers × **4** views = 16 per task. Durations are reported in minutes: each clip is a single clean execution of one manipulation or locomotion primitive rather than a long-horizon episode. Team work and box transfer are two-performer interaction tasks; all others are single-performer.

Task	Clips	Minutes
Packing	16	106.3
Arranging household items	16	59.0
Ironing clothes	16	52.0
Cloth cutting	16	43.9
Clothes folding	16	41.7
Team work (two performers)	16	37.9
Putting clothes on	16	37.5
Chopping and cooking	16	27.8
Stacking books	16	21.8
Sweeping	16	21.5
Sweep and collect dust	16	18.5
Arranging bottles	16	18.5
Canister vacuum cleaner	16	18.3
Vacuum cleaner	16	18.0
Flat-mop sweeping	16	17.9
Mopping	16	17.5
Steps climbing	16	14.2
Box transfer (two performers)	16	7.7
Total	**288**	**580.0**

Table 1. Headline statistics for Bumblebee across three synchronized modalities and household/factory domains. Values marked ‡ are measured on a random subset of household clips; all others are measured on the full released corpus after final QC. Detailed breakdowns are in Sections 3–5 and Appendices A–D.

Egocentric Stream		**Tri-Camera Stream**	
Household hours	**11,778**	Curated hours	**75**
Household clips	**29,247**	Clips (**100**×**3**)	**300**
Household participants	**1,802**	Viewpoints	**head + 2 wrists**
Factory hours	**2,544**		
Factory clips	**1,283**		
Factory participants	**482**		
Motion Capture (marker-based)		**Curation & Annotations**	
Curated hours	**9.7**	Metadata versioning	**shipped**
Approved clips (**72**×**4**)	**288**	Task granularities	**3**
Task categories	**18**	Skeletal ground truth	**metric 3D**
Performers	**4**	Clips per task category	**16**
Format (Egocentric)		**Format (MoCap)**	
Native resolution	**FHD–4K**	Capture method	**optical marker**
Native frame rate	**30–120 fps**	Skeletal output	**3D joint traj.**
Audio	**synchronized**	Reference video	**synchronized**
Mean clip duration	**24** min	Timecode logging	**per take**
Hand Visibility & Diversity		**Integrity (All Streams)**	
Hand-visible frames‡	**68%**	Cut-free clips‡	**43%**
Both-hands visibility‡ (>**20%**)	**45%**	Undocumented cuts	**0**
Household domains / tasks	**9** / **50**	Duplicates (all modalities)	**0**
Factory families / tasks	**10** / **52**	Modal sync accuracy	**frame-exact**
Mean illumination†‡	**0.44**		
Provenance			
Consent & commercial license	**100%**	Datasheet	shipped
Identifiable non-consenting	**0**	Chain-of-custody documentation	shipped
		Independent-audit access	under NDA
		License	commercial

† *Mean illumination is the mean per-frame luminance of egocentric video, normalized to the* $[0, 1]$ *range.* ‡ *Measured on a random subset of household clips, not on the full corpus.*

Table 2. Composition by household activity domain, measured on the released corpus. Percentages are shares of total recorded hours.

Activity domain	Clips	Hours	% hours
Cleaning	12,021	4,248	36%
Organization	6,697	2,387	20%
Cooking & food prep	2,419	1,820	15%
Laundry management	4,488	1,355	12%
Daily routines	709	612	5%
Assistance & navigation	1,933	538	5%
Preparation & occasions	445	381	3%
Hosting	305	258	2%
Home maintenance	230	179	2%
Total	**29,247**	**11,778**	**100%**

Table 3. Composition by factory operation family, measured on the released corpus. Families group tasks by required manipulation skill. Percentages are shares of total factory recorded hours.

Operation family	Tasks	Hours	% hours
Component insertion & assembly	5	560	22%
Inspection & quality verification	9	549	22%
Precision soldering & joining	4	546	21%
Electrical test & debug	8	288	11%
Packaging & kitting	5	160	6%
Material processing & forming	6	155	6%
Machine tending & material handling	5	120	5%
Surface preparation & cleaning	4	65	3%
Marking & labeling	3	64	3%
Line supervision & unclassified	3	37	1%
Total	**52**	**2,544**	**100%**