BARISTA: A Multi-Task Egocentric Benchmark for Compositional Visual Understanding

Patrick Knab [*1,2] **Orgest Xhelili** [*2] **Inis Buzi** [2] **Drago Andres Guggiana Nilo** [2] **Mohd Saquib Khan** [2]
Lorenz Kolb [2] **Manuel Scherzer** [2] **Kerem Yildirir** [2] **Christian Bartelt** [1] **Philipp J. Schubert** [2]

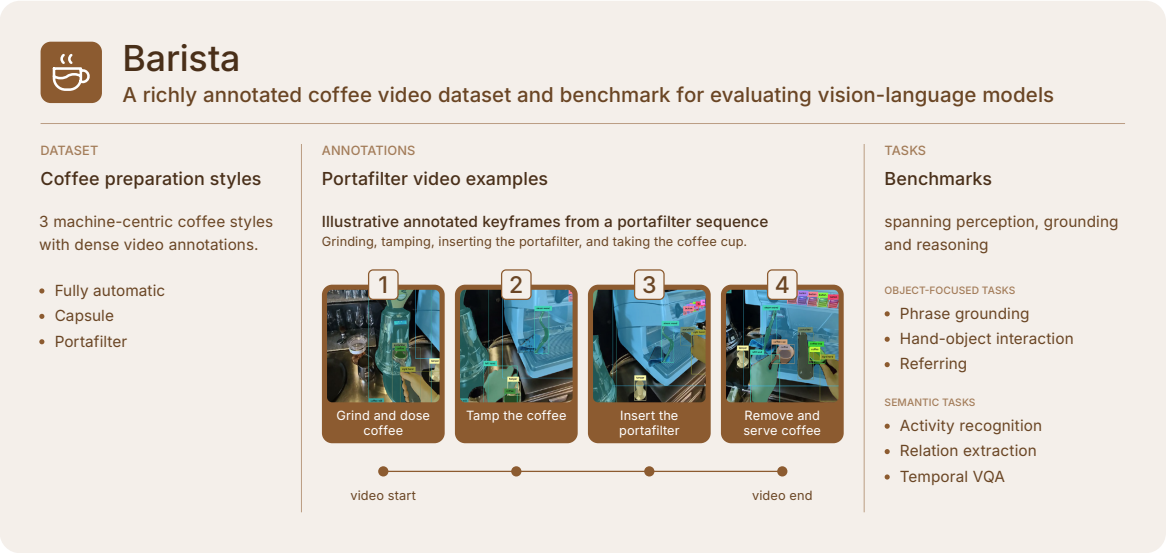

Figure 1. BARISTA overview. BARISTA is an egocentric coffee-preparation dataset and benchmark spanning three preparation styles, dense task-relevant annotations, and object-focused and semantic evaluation tasks derived from a shared scene graph.

Abstract

Scene understanding is central to general physical intelligence, and video is a primary modality for capturing both state and temporal dynamics of a scene. Yet understanding physical processes remains difficult, as models must combine object localization, hand-object interactions, relational parsing, temporal reasoning, and step-level procedural inference. Existing benchmarks usually evaluate these capabilities separately, limiting diagnosis of *why* models fail on procedural tasks. We introduce **BARISTA**, a densely annotated egocentric dataset and benchmark of 185 real-world coffee-preparation videos covering fully automatic, portafilter-based, and capsule-based workflows. BARISTA provides verified per-frame scene graphs linking persistent object identities to masks, tracks, boxes, attributes, typed relations, hand-object interactions, activities, and process steps. From these graphs, we derive zero-shot language-based tasks spanning phrase grounding, hand-object interaction recognition, referring, activity recognition, relation extraction, and temporal visual question answering. Experiments reveal strong variation across task families and no consistently dominant model family, positioning BARISTA as a challenging diagnostic benchmark for procedural video understanding. Code and dataset available at https://huggingface.co/datasets/ramblr/BARISTA.

[1]Technical University of Clausthal [2]Ramblr.ai Research. Correspondence to: Philipp J. Schubert <philipp@ramblr.ai>, Patrick Knab <patrick.knab@tu-clausthal.de>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

1. Introduction

Vision-language models (VLMs) increasingly serve as the perception backbones of agentic and embodied systems

operating in the physical world (Driess et al., 2023; Kim et al., 2024; Gao et al., 2024; Assran et al., 2025; Team et al., 2023; Ahn et al., 2022). Such systems require a coordinated set of capabilities: localization of task-relevant objects and regions, interpreting interactions with the environment, recognizing ongoing activities in context, and reasoning about how the scene evolves over time (Team et al., 2023; Intelligence et al., 2025; Knab et al., 2025; Cho et al., 2025). These abilities span multiple levels of abstraction, from masks and boxes to typed relations and process-level reasoning, and are difficult to assess when each capability is benchmarked on a different dataset under its own assumptions and evaluation protocol. This points to a core benchmark-design requirement: evaluations must support controlled, fine-grained analysis of failure across levels, not merely report aggregate scores. Most benchmarks isolate a single aspect of compositional understanding, such as activity recognition, long-video question answering, or object-centric perception (Mangalam et al., 2023; Li et al., 2024; Fu et al., 2025; Salehi et al., 2024). As a result, failures are hard to interpret: errors on process-level tasks may stem from weak localization, interaction understanding, relational reasoning, or temporal modeling, without a way to disentangle them. This is particularly problematic given recent evidence that VLMs remain brittle even on basic spatial cognition (Ahn et al., 2022; Khemlani et al., 2025; Fu et al., 2024; Prasse et al., 2026). Existing egocentric and instructional video datasets offer a path to tackling this—providing dense spatial annotations, activity labels, or relational structure—but each typically covers only part of the required supervision. Because their domains, label spaces, and evaluation setups differ, they provide limited common ground to trace a high-level reasoning failure back to a concrete perceptual gap.

We introduce BARISTA, a densely annotated egocentric dataset and a unified benchmark of 185 real-world coffee-preparation videos designed to address this gap. The technical core of BARISTA is a per-frame co-registered scene graph. Each annotated frame includes instance masks, bounding boxes, per-instance attributes, and directed typed relations anchored to persistent object identities, together with activity and process-step labels. We deliberately trade duration for annotation density and traceability: the 4.4 hours of BARISTA carry 3.61M instance masks, 4.33M attributes, and 2.48M typed relations, all co-registered through persistent identities. This structure supports high-quality multi-task annotation at practical scale and naturally enables query-driven VLM evaluation. More importantly, it makes BARISTA explicitly decomposable: object-, interaction-, relation-, and process-level tasks are defined over the same underlying evidence, enabling controlled comparisons and hypothesis testing about which lower-level failures drive higher-level errors. Coffee preparation serves here as a compact yet expressive procedural testbed: it is structured enough to require process reasoning, varied enough across machine types to expose differences in tools, interfaces, interaction patterns, and step order, and familiar enough that failure modes are easy to inspect.

Figure 1 provides an overview of BARISTA, illustrating the covered coffee-preparation styles, dense annotation layers, and the benchmark tasks. For object-based perception, we provide annotations for phrase grounding, hand-object interaction recognition, segmentation masks, and referring expression generation; for semantic scene understanding, we focus on activity recognition, relation extraction, and temporal visual question answering. The unified benchmark evaluates language-facing tasks under a shared zero-shot protocol, while the released annotations additionally support downstream training and fine-tuning of specialist perception models. Because all language-facing benchmark items are derived deterministically from this verified graph, failures on process-level tasks can be traced back on the same frames to the localization, interaction, and relation annotations probed by the other tasks. We benchmark four VLM families — Qwen (Yang et al., 2025a), Gemma (Team et al., 2024), GPT (Achiam et al., 2023), and Gemini (Team et al., 2023) — across different model sizes, finding that no model performs uniformly best across the different tasks. Notably, spatial grounding performance varies far more across models than semantic recognition, and the capabilities underlying localization-heavy tasks are largely orthogonal to those driving temporal QA.

Our main contributions are as follows:

- We introduce BARISTA, a densely annotated egocentric dataset of 185 coffee-preparation videos built around a per-frame co-registered scene graph linking masks, boxes, attributes, relations, hand-object interactions, and procedural steps via persistent object identities.
- We propose a semi-automated annotation pipeline and a graph-based task-generation procedure for dense spatial, relational, interaction, and language-derived supervision.
- We define a unified, decomposable benchmark with standardized zero-shot prompting and task-specific metrics, and use it to show that current frontier and open-weight VLMs exhibit *task-orthogonal* capability profiles.

2. Related work

Egocentric and instructional video datasets. Large-scale egocentric resources such as EPIC-KITCHENS (Damen et al., 2018), Ego4D (Grauman et al., 2022), Ego-Exo4D (Grauman et al., 2024), and the long-form Ego-Life (Yang et al., 2025b) have substantially advanced first-

Table 1. **Comparison of selected related video datasets.** *sparse*: objects annotated only at movement events. *contact*: hand-object contact only. *fixture*: object-fixture associations from 3D lifting. *typed*: semantic object-object or hand-object relation labels beyond boxes/masks.

Dataset	View	Dataset Size	Instance Masks	Bounding Boxes	Object Attributes	Typed Relations	Action Labels	Temporal Tracks
EPIC-KITCHENS (Damen et al., 2018)	Ego	100h					✓	
VISOR (EPIC) (Darkhalil et al., 2022)	Ego	36h	✓	✓		contact		✓
HD-EPIC (Perrett et al., 2025)	Ego	41h	sparse	✓		fixture	✓	✓
Ego4D (Grauman et al., 2022)	Ego	3000h		✓			✓	
Ego-Exo4D (Grauman et al., 2024)	Ego+Exo	1286h		✓			✓	
Assembly101 (Sener et al., 2022)	Ego+Exo	513h					✓	
EPFL-SmartKit. (Bonnetto et al., 2025)	Ego+Exo	29.7h					✓	
PVSG (Yang et al., 2023a)	Ego+Exo	9h	✓	✓		✓		✓
BARISTA (Ours)	Ego	4.4h	✓	✓	✓	✓	✓	✓

person video understanding, primarily through narration-derived action labels and temporal boundaries rather than per-frame structured annotations. VISOR (Darkhalil et al., 2022) extends EPIC-KITCHENS with 272K manual pixel masks for hands and active objects across 257 classes, 9.9M interpolated dense masks, and 67K hand-object contact-state relations, but does not annotate typed semantic relations between objects, per-instance attribute vocabularies, or activity segments.

HD-EPIC (Perrett et al., 2025) is the closest recent point of comparison, providing 41.3 hours of unscripted kitchen recordings with dense hand masks, sparse object masks, 3D digital twins, gaze, audio, and a 26K-question VQA benchmark. However, object masks are temporally sparse, and the only explicitly annotated relation type connects objects to kitchen fixtures via 3D lifting; other inter-object and hand-object relations must be recovered from narration parsing. Egocentric procedural datasets provide complementary supervision. Assembly101 (Sener et al., 2022) and HOI4D (Liu et al., 2022) offer dense 3D hand pose and multi-granularity activity segments, with HOI4D adding panoptic segmentation and 6-DoF object pose. HoloAssist (Wang et al., 2023) and EPFL-Smart-Kitchen-30 (Bonnetto et al., 2025) capture rich multi-modal signals (gaze, dialogue, 3D body, and hand pose) for interactive assistance and fine-grained cooking. Instructional benchmarks such as YouCook2 (Zhou et al., 2018), COIN (Tang et al., 2019), and EgoProceL (Bansal et al., 2022) provide temporal step boundaries at scale but lack per-frame spatial annotations. Scene-graph datasets such as Action Genome (Ji et al., 2020), PVSG (Yang et al., 2023a), and AGQA (Grunde-McLaughlin et al., 2021) offer relational supervision but outside of procedural egocentric settings.

VLM benchmarking. Recent video-VLM benchmarks evaluate complementary aspects of video understanding: EgoSchema (Mangalam et al., 2023) and LongVideoBench (Wu et al., 2024) test long-form temporal reasoning over minutes to hours, MVBench (Li et al., 2024) and Video-MME (Fu et al., 2025) provide broad multi-task or multi-domain coverage, and ActionAtlas (Salehi et al., 2024) and FAVOR-Bench (Tu et al., 2025) target fine-grained action or motion discrimination. Together, they span a diverse evaluation landscape from temporal localization and action discrimination to broad multi-domain video comprehension. However, these benchmarks typically evaluate final-answer accuracy rather than grounding each question in a verified per-frame scene graph.

Positioning. Table 1 summarizes the datasets most directly comparable to BARISTA. BARISTA is smaller in scale (185 videos, $\sim$469K annotated frames) but provides dense per-frame co-registered supervision: 3.61M instance masks across 46 object categories, 2,424 activity segments over 108 classes. Every benchmark task is derived deterministically from this verified scene graph, and each item is grounded in verified instance masks, relations, and activity segments, so that model failures can be diagnosed at the spatial and interaction level. Unlike broad video leaderboards that primarily aggregate end-task scores, BARISTA is designed as a controlled, decomposable benchmark whose task families help to expose how lower-level perceptual errors propagate into higher-level reasoning failures.

3. BARISTA

3.1. Dataset creation

Coffee preparation tasks. Coffee preparation is a compact but expressive procedural domain. Different preparation styles share the same high-level goal while varying substantially in required tools, visible machine state, hand usage, step order, and interaction duration. BARISTA focuses on

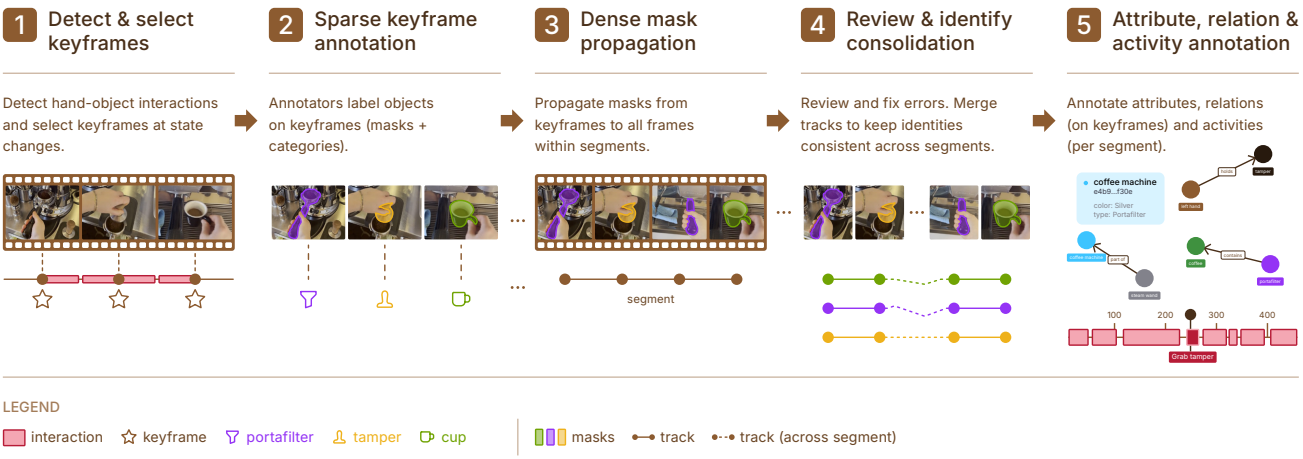

Figure 2. **BARISTA annotation pipeline.** Egocentric coffee-preparation videos are reduced to interaction segments and keyframes, sparsely annotated with object masks and categories, densely propagated across frames, reviewed for identity consistency, and enriched with attributes, typed relations, activities, and procedural step labels. This produces a per-frame co-registered scene graph linking spatial, relational, interaction, and temporal supervision.

three representative preparation types:

- *Fully automatic machines*, which emphasize short but visually diverse interactions such as button pressing, milk handling, cup placement, and machine-state feedback.
- *Capsule machines*, which involve inserting a capsule, opening and closing the lever, and automated extraction with substantial variation in the machine form factor.
- *Portafilter-based espresso preparation*, which captures a substantially more manual workflow, including dosing, tamping, attaching the portafilter, and starting extraction.

Together, these procedures span both fast interface-driven actions and longer manipulation-heavy sequences, making them useful for studying localization, hand-object interaction, relation extraction, activity recognition, and process reasoning in one domain.

Data collection procedure. All 185 videos were recorded in-house by people affiliated with the research group in controlled real-world coffee-preparation setups. No external participants or crowdworkers were involved in data capture. Recording devices include smartphones and smart glasses (Apple Vision Pro, RayBan Meta 3, RayBan Wayfarer), so that the same procedural domain is captured across heterogeneous egocentric hardware. The recorders varied tools, machines, coffee variants, and optional steps such as grinding or milk preparation, which introduces natural procedural variation without changing the underlying domain. The dataset is released under a CC-BY-NC 4.0 license and hosted on Hugging Face (Appendix A.4).

3.2. Annotation pipeline

Producing temporally consistent supervision for egocentric manipulation videos is challenging because hands frequently occlude objects, camera motion is substantial, and meaningful actions can be brief. To scale BARISTA while preserving quality, we build a per-frame co-registered scene graph with a semi-automated pipeline that alternates model-assisted bootstrapping with targeted human verification (Chen et al., 2026; 2025). Annotations were performed by trained annotators affiliated with the authors' research institution. Figure 2 visualizes the annotation pipeline described in the following.

Step 1: Interaction detection and keyframe selection. We first run a CaRe-Ego model (Su et al., 2026), finetuned on an internal egocentric hand-object interaction dataset used only for annotation bootstrapping, on every frame to produce per-frame interaction confidence scores. We extract interaction segments via temporal smoothing and morphological filtering of the raw signal (median filtering and binarization, connected-components extraction, removal of short segments, and merging of close gaps). Within each interaction segment, we select its temporal center as a keyframe while maintaining a maximum inter-keyframe distance of 500 frames.

Step 2: Sparse keyframe annotation. Annotators label all scene-relevant objects on the selected keyframes using a custom interface with interactive mask suggestions from SAM 2 (Ravi et al., 2025). They can accept a proposed mask or refine it with brush and polygon tools. Each instance receives a mask and category label and previously annotated objects are re-presented within the same segment to encourage identity reuse.

Step 3: Dense mask propagation. We expand the sparse

keyframe annotations into dense per-frame masks using the video propagation capabilities of SAM 2 (Ravi et al., 2025) and SAM 3 (Meta FAIR, 2025). For each propagation segment, the model is conditioned on all annotated keyframe masks simultaneously and propagates bidirectionally, so that intermediate frames benefit from reference appearances on both sides. This multi-frame conditioning provides greater robustness to occlusion, motion blur, and appearance change compared to single-frame propagation. Additionally, adjacent segments are propagated with shared temporal overlap, linking object identities across segment boundaries based on mask agreement.

Step 4: Review and identity consolidation. Annotators review the propagated annotations in a correction interface to fix low-quality masks, missing masks after occlusion, or identity inconsistencies. They correct only sparse frames and repropagate within the problematic segments, keeping review efficient. Annotators then consolidate object identities across segments: when the same physical object reappears after occlusion or across neighboring segments, its per-frame sequences are merged into a single consistent identity.

Step 5: Attribute, relation, and activity annotation. We complete the scene graph with three segment-level layers (object attributes, directed relations, and activities). Annotation is guided by a controlled domain vocabulary that is initialized before annotation and extended only through curator-approved additions when genuinely new entities appear. Object-level attributes are assigned over each instance's active frame range. Directed typed relations are labeled on representative frames and propagated across a given segment, distinguishing spatial relations between objects (e.g. on, inside, part of) from hand-object actions (e.g. holds, presses, touches). Lastly, activities are annotated at two granularities: fine-grained verb+noun labels (e.g. place cup) and coarser procedural steps reflecting the high-level process structure.

3.3. Dataset analysis

BARISTA comprises 185 videos and 469,201 annotated frames with a dense graph-structured annotation layer: 3.61M instance masks with bounding boxes across 46 object categories, 4.33M per-instance attribute annotations (color, type, state), 2.48M typed relations (12 predicates covering spatial arrangement and hand-object interactions), 2,424 fine-grained activity segments spanning 108 classes, and 1,305 process steps across 13 classes. Videos are exported at 30 FPS across heterogeneous devices and resolutions ranging from 1280×720 to 1920×1080. Annotated frames contain on average 7.63 objects (median 7, up to 24) and 5.3 typed relations per frame. A detailed distributional analysis—including per-category frequencies, activity class counts, relation predicate distributions, and duration statistics—is provided in subsection A.6. All reported benchmark experiments are conducted in a zero-shot setting without any training phase, using fixed evaluation subsets sampled from the full dataset. To facilitate future training experiments, the BARISTA release includes a video-level train/test split (80/20), stratified by preparation type, with 148 and 37 videos, respectively.

4. BARISTA benchmark

4.1. Unified evaluation protocol

The BARISTA benchmark defines a suite of zero-shot tasks that probe complementary aspects of egocentric procedural understanding. All tasks are derived deterministically from the verified scene graph described in section 3, ensuring that each evaluation example is traceable to the same underlying annotations. Depending on the task, examples use either single-frame input (phrase grounding, hand-object interaction recognition, referring expression generation, and relation extraction) or multi-frame clips (activity recognition and temporal visual question answering). Taken together, the task families form a layered evaluation stack, from region-level localization to short-horizon process reasoning. For each task, annotated frames or clips are paired with task-specific ground truth, including bounding boxes, interaction labels, relation tuples, or action labels. At evaluation time, each example is rendered into a multimodal prompt, and the model's response is scored against the ground truth with a task-appropriate metric.

Sampling and metrics. To keep the evaluation cost manageable while preserving task diversity, we use fixed evaluation subsets in the main benchmark: 1000 samples for single-frame tasks, 500 samples for activity recognition with 4 frames per example, and 300 samples for temporal VQA with 6 frames per example. Because all samples are derived from dense annotations through deterministic sampling with configurable strides and per-class limits, users can scale each benchmark to larger evaluation sets, re-sample it with different criteria, or run standardized ablations over shared subsets. This scalability also supports fine-grained error analysis, including how errors propagate across tasks: users can sample additional instances from critical video segments and inspect model behavior in regions of interest. Evaluation uses task-appropriate metrics, including IoU-based detection metrics for localization, set-matching F1 for structured outputs, LLM-as-judge G-Eval scoring for open-ended textual responses, and classification accuracy for recognition tasks. More details can be found in A.

Models. We evaluate general-purpose VLMs under a uniform zero-shot protocol. Our main evaluation covers two closed models, Gemini 3 Flash (Team et al., 2023) and

GPT-5.4 mini (Achiam et al., 2023), and three open-weight models, Qwen3.5-27B, Qwen3.5-9B (Yang et al., 2025a), and Gemma 4-31B (Team et al., 2024). Together, they span proprietary and open-weight VLMs across different families and scales, balancing state-of-the-art reasoning capability with practical cost and inference efficiency.

Each task specifies a task-specific system prompt and a user-prompt template, and these prompts are shared across models with minimal provider-specific adaptations (e.g. bounding-box format). For closed-source models, we set the thinking effort to low and query with temperature 0.7, except for GPT-5.4 mini, which does not allow for custom sampling. We run the evaluations three times and report the mean and standard deviation. Two specialist baselines complement the VLMs: CaRe-Ego (Su et al., 2026) for egocentric hand-object interaction, and SAM 3 (Meta FAIR, 2025) for open-vocabulary grounding. Each benchmark example is released together with the annotations so that the evaluation is reproducible and provider-agnostic, covering both hosted APIs and open-weight checkpoints served through vLLM (Kwon et al., 2023).

4.2. Task families with benchmarks

4.2.1. Object-focused tasks

Phrase grounding (Liu et al., 2024; You et al., 2024) evaluates whether a model can link a short natural-language phrase, such as *"the red button on the automatic coffee machine"* to the correct object in a coffee-making scene. Given a single frame, the model must predict a bounding box for the referenced object. Phrases are constructed from verified category, color, and spatial-relation annotations, selecting the richest available description for each category, which makes them context-dependent and rarely reducible to a simple category label. Performance is measured using COCO-style (Lin et al., 2014) mean average precision (mAP): mAP@50 measures detection quality at a loose IoU threshold of 0.5 (correct region), while mAP averages over IoU thresholds from 0.5 to 0.95 and rewards tight box localization. Full size-stratified results are reported in Appendix A.2.1. We compare zero-shot VLMs with SAM 3 as an open-vocabulary grounding specialist on the same frozen evaluation set.

As shown in Table 2, Gemini 3 Flash and Qwen3.5-27B lead overall. SAM 3 is competitive but remains below the strongest VLMs, whereas GPT-5.4 mini performs poorly in this setting—it is included as a general-purpose VLM reference but is not primarily optimized for region-conditioned visual grounding. A per-size-bin breakdown (Appendix A.2.1) shows that performance drops substantially on medium-sized objects, indicating that object scale remains a key bottleneck.

Table 2. **Phrase grounding.** COCO-style mAP on 1000 frozen single-frame examples.

Model	mAP@50	mAP
SAM 3	.393	.362
Gemini 3 Flash	**.697** ± .017	**.534** ± .008
GPT-5.4 mini	.163 ± .022	.068 ± .014
Gemma 4-31B	.634 ± .005	.432 ± .004
Qwen3.5-27B	.680 ± .016	.518 ± .005
Qwen3.5-9B	.559 ± .027	.350 ± .020

Egocentric hand-object interaction recognition (Su et al., 2026; Zhang et al., 2022) requires the model to detect and characterize all hand-object interactions visible in a single egocentric frame. For each interaction, the ground truth specifies the manipulated object, the involved hand(s), and a hand-type label (left, right, or both). The model must localize both the object and the hand regions via bounding boxes and classify the type of hand interaction (whether the left hand, right hand, or both hands are involved). The predicted interactions are greedily matched to the ground-truth annotations by object-box IoU at a 0.5 threshold. On matched pairs, we report interaction-level F1 and mean hand and object IoU. Precision, recall, no-detection rate, and hand-type classification accuracy are reported in Appendix A.2.2. We evaluate CaRe-Ego as a baseline specialist model.

Table 3. **Egocentric hand-object interaction.** Interaction F1 and localization IoU on 1000 frozen frames (IoU≥0.5 matching).

Model	Int. F1	Hand IoU	Obj. IoU
CaRe-Ego	.490	.703	.839
Gemini 3 Flash	.749 ± .008	**.744** ± .007	**.892** ± .001
GPT-5.4 mini	.193 ± .014	.483 ± .007	.644 ± .008
Gemma 4-31B	**.758** ± .003	.707 ± .001	.857 ± .001
Qwen3.5-27B	.753 ± .007	.706 ± .004	.864 ± .002
Qwen3.5-9B	.642 ± .008	.640 ± .003	.787 ± .005

Table 3 shows that the strongest VLMs outperform CaRe-Ego on interaction F1, while CaRe-Ego remains competitive in localization, especially for object boxes. Among the VLMs, Gemma 4-31B, Qwen3.5-27B, and Gemini 3 Flash are tied on interaction F1 (.749–.758), while Gemini 3 Flash produces the most accurate hand and object boxes.

Referring / region description (You et al., 2024; Guo et al., 2024) is the inverse of phrase grounding: instead of mapping language to a region, the model must describe a given region in context. Each example consists of a frame and a specific bounding box, and the model must generate a concise description that identifies the marked object through its category, visual attributes, and spatial or functional relations to surrounding objects. Ground-truth references are

assembled deterministically from the verified structured annotations, listing the object's category, attributes, direct relations, and context about related objects. A separate LLM judge (Gemini 2.5 Flash) scores each prediction against this structured reference using G-Eval (Liu et al., 2023) under a single Correctness criterion on a 0–10 rubric (rescaled to $[0, 1]$). Because LLM-based evaluation is not calibrated, absolute scores are best interpreted for relative model comparison rather than as ground-truth quality estimates. The rubric and full prompts are detailed in Appendix A.2.3.

Table 4 shows that Gemini 3 Flash performs best, with Gemma 4-31B the strongest open-weight model, followed by Qwen3.5-27B and Qwen3.5-9B. GPT-5.4 mini lags substantially behind, mirroring its weak performance on phrase grounding due to its limitations in spatially grounded understanding. Model rankings are stable when re-scored by a cross-family judge ($\tau = 1.0$; Appendix A.7), confirming the ordering reflects task performance rather than judge-specific bias.

Table 4. **Referring expression generation.** G-Eval correctness on 1000 frozen examples.

Model	Correct.
Gemini 3 Flash	**.760** ± .003
GPT-5.4 mini	.308 ± .006
Gemma 4-31B	.667 ± .004
Qwen3.5-27B	.622 ± .010
Qwen3.5-9B	.576 ± .009

4.2.2. Semantic tasks

Activity recognition (Darkhalil et al., 2022; Feichtenhofer et al., 2019) evaluates whether a model can identify a fine-grained manipulation activity from a short frame sequence. Given four frames sampled uniformly from an annotated activity segment, the model must select the correct activity label from K=25 multiple-choice options drawn from the full activity vocabulary. We choose K=25 to make the task non-trivial while keeping the option set interpretable (constructions detailed in Appendix A.3.1).

The results in Table 5 show that Qwen3.5-27B and Gemini 3 Flash are effectively tied at the top (.778 vs .765). The Qwen models show particular strength on this task, with Qwen3.5-9B outperforming both Gemma 4-31B and GPT-5.4 mini despite its smaller scale.

Table 5. **Activity recognition.** 25-way MCQ accuracy on 500 clips with 4 frames each.

Model	Acc.
Gemini 3 Flash	.765 ± .015
GPT-5.4 mini	.631 ± .007
Gemma 4-31B	.684 ± .003
Qwen3.5-27B	**.778** ± .005
Qwen3.5-9B	.716 ± .014

Relation extraction evaluates whether a model can recover the relational structure of a scene, moving beyond object recognition to structured scene understanding (Krishna et al., 2017). The model receives a single frame with a set-of-mark (Yang et al., 2023b) overlay labeling each annotated object with a numerical identifier, a legend mapping identifiers to categories, and the constrained vocabulary of valid relation types and values. Given this input, the model must predict all directed relation tuples (s, t, r, v)—source object, target object, relation type, and value—that hold between the identified objects. Predictions are evaluated against the verified ground-truth set via exact multiset matching: a predicted tuple counts as correct only if it matches a ground-truth tuple on all four fields.

Table 6. **Relation extraction.** Micro F1 under exact (s, t, r, v) multiset matching on 1000 frozen frames.

Model	Overall	human_actions	position
Gemini 3 Flash	.675 ± .008	**.675** ± .004	.710 ± .007
GPT-5.4 mini	.565 ± .008	.534 ± .011	.569 ± .008
Gemma 4-31B	**.751** ± .002	.673 ± .012	**.758** ± .001
Qwen3.5-27B	.745 ± .005	.626 ± .006	.756 ± .004
Qwen3.5-9B	.613 ± .004	.573 ± .010	.630 ± .005

Because object identities and categories are provided, this task isolates relation prediction rather than object detection. We report overall and per relation type F1 and full metrics in A.3.2. In Table 6, Gemma 4-31B achieves the best overall F1, narrowly ahead of Qwen3.5-27B.

Temporal visual question answering evaluates open-ended reasoning (Yu et al., 2019) over short temporal windows rather than single images. We detect temporal changes between clip endpoints from structured annotations—attribute transitions, hand-contact sequences, spatial-relation shifts, and activity-label boundaries—and prompt Gemini 2.5 Flash Lite to convert each detected change into a natural-language question-answer pair. The resulting pairs span four categories: *state change*, *action transition*, *relation change*, and *activity progression*. At inference time, the model receives six frames from a four-second window and one question and must produce a free-form answer. A separate LLM judge (Gemini 2.5 Flash) scores each response with G-Eval (Liu et al., 2023) on Accuracy (factual correctness) and Completeness (coverage of reference changes). As in referring, the absolute G-Eval scores reflect judge-specific calibration and are most informative for ranking models rather than measuring absolute answer quality. Rankings are preserved under a cross-family GPT-5.4 mini judge on mean score and Completeness ($\tau = 1.0$), while a single adjacent Accuracy swap occurs between Gemini 3 Flash and GPT-5.4 mini (Appendix A.7). The complete generation and LLM judging details are in Appendix A.3.3.

Table 7 shows that Qwen3.5-27B has the highest mean and accuracy scores, while being tied with GPT-5.4 mini on completeness. GPT-5.4 mini performs well, ranking second on mean score despite weaker performance on other tasks. Per-category results (Appendix A.3.3) show that relation-

Table 7. **Temporal VQA.** LLM-judged accuracy and completeness (G-Eval, $[0, 1]$) on 300 clips (6 frames, 4 s).

Model	Accuracy	Completeness	Mean
Gemini 3 Flash	$0.696 \pm .015$	$0.640 \pm .008$	$0.665 \pm .009$
GPT-5.4 mini	$0.701 \pm .007$	$0.675 \pm .008$	$0.686 \pm .003$
Gemma 4-31B	$0.672 \pm .008$	$0.629 \pm .009$	$0.649 \pm .007$
Qwen3.5-27B	$\mathbf{0.726} \pm .004$	$\mathbf{0.680} \pm .005$	$\mathbf{0.701} \pm .006$
Qwen3.5-9B	$0.674 \pm .016$	$0.641 \pm .017$	$0.655 \pm .017$

change questions are answered most reliably, whereas action transitions remain the most difficult category.

5. Discussion and conclusion

Key findings. BARISTA shows that holistic video understanding does not reduce to a single notion of "multimodal capability": models that excel on one part of the pipeline may still perform poorly on another. Across the benchmark, no model family dominates every task. Gemini 3 Flash is strongest on object-focused tasks that depend on precise visual grounding—phrase grounding, referring expression generation, and accurate hand/object boxes for HOI—whereas the strongest open-weight models lead on structured prediction, with Gemma 4-31B achieving the best relation-extraction F1 and Qwen3.5-27B the best temporal VQA and activity-recognition performance. This pattern suggests that BARISTA separates distinct capabilities: localization, interaction recognition, relational parsing, activity recognition, and short-horizon temporal reasoning.

A second consistent finding is that spatial grounding remains the main bottleneck. Performance varies most on tasks requiring reliable region-level alignment, especially phrase grounding and HOI. Even strong models degrade markedly on medium-sized objects: mAP drops by more than half between the large and medium size bins for every model, and the performance spread across models on grounding (mAP range 0.07–0.53) is far wider than on semantic tasks such as activity recognition (accuracy range 0.63–0.78) or temporal VQA (mean range 0.65–0.70). Current VLMs often capture the stage of a process before they can consistently localize the objects and interactions that explain it. This asymmetry has direct implications for embodied deployment: a system that recognizes activities but cannot reliably locate manipulated objects will struggle to provide actionable feedback in a physical workflow.

A third finding is that BARISTA exposes a meaningful internal structure within "procedural understanding". Within temporal VQA, relation changes are answered most reliably, whereas action transitions remain the hardest category, suggesting that models find static before/after differences easier than fine-grained motion or transition dynamics from sparse frame sequences. This heterogeneity extends to the benchmark level: a post-hoc cross-task rank analysis (Appendix A.8) shows that grounding, HOI, relations, and referring form a coherent cluster whose model rankings agree closely (Kendall $\tau \geq 0.6$ on 5 of 6 pairs), while temporal VQA is decoupled from this cluster ($\tau \in [-0.4, 0.0]$) and correlates only weakly with activity recognition (τ=0.4). In other words, the capabilities that make a model good at localization and structured spatial reasoning are largely orthogonal to those that drive temporal QA performance. Per-machine-type stratified results (Appendix A.9) further show that the three preparation styles expose distinct performance profiles across tasks, confirming that procedural diversity is a meaningful evaluation axis.

Limitations. BARISTA is intentionally narrow: it focuses on a single procedural domain, indoor coffee preparation, from an egocentric viewpoint. We therefore do not claim that performance on BARISTA fully characterizes general physical-world understanding. Our annotations are dense over task-relevant entities and interaction segments rather than exhaustive over every visible object in every frame, reflecting the benchmark's emphasis on compositional process understanding rather than open-world scene coverage. The activity-recognition task is formulated as multiple-choice classification with the correct answer present among the candidates, making the evaluation controlled and comparable, but likely overestimating performance relative to open-vocabulary settings. Parts of the benchmark depend on LLM-as-a-judge scoring for free-form outputs, even though a cross-family judge swap (Appendix A.7) preserves model rankings ($\tau \geq 0.8$ on all metrics, $\tau = 1.0$ on three of four), suggesting that reported orderings reflect genuine performance differences rather than judge-specific artifacts. Finally, several evaluations use frozen subsets to keep benchmarking cost manageable, although the released annotations and sampling scripts make future expansion straightforward.

Dataset utility beyond benchmarking. BARISTA is also designed to be useful as a training and analysis infrastructure. Temporally consistent masks and tracks support segmentation and tracking; phrase-grounding pairs and boxes support localization; and relation tuples, HOI labels, and activity segments provide structured supervision for fine-tuning specialized perception modules. Because all labels are linked through persistent object identities, the dataset can further support work on temporal scene graphs, interaction forecasting, process monitoring, failure analysis, and compositional debugging of vision-language systems. More broadly, BARISTA offers a way to study how low-level perceptual errors propagate into higher-level reasoning failures within one annotation framework.

Conclusion. BARISTA advances compositional visual understanding by combining dense, task-relevant annotation

with unified multi-task evaluation in a realistic egocentric procedural setting. Its central contribution is diagnostic coherence and decomposability: the benchmark can trace a process-level failure back to localization, interaction recognition, relational structure, or temporal inference on the same underlying evidence. Our experiments show that current VLMs remain uneven across these capabilities, with comparatively strong semantic recognition still falling short of consistently grounded procedural understanding. By releasing videos, annotations, prompts, splits, and evaluation code, we provide a compact but extensible testbed for measuring not only whether procedural vision-language systems succeed, but where and why they fail.

Ethics Statement. The dataset was recorded by the authors in controlled lab environments using egocentric cameras, primarily capturing the recorder's hands and workspace during the coffee-making activity. In rare cases, faces are partially visible in reflective surfaces, for which all affected individuals provided informed consent for publication. All annotators were affiliated with the research institution and compensated appropriately. We only evaluate pre-trained vision-language models and release no fine-tuned models, limiting potential for misuse. The dataset and code are released with explicit licenses and documentation to ensure responsible use and reproducibility.

6. Acknowledgments

This work was partially funded by the Bavarian State Ministry for Economic Affairs, Regional Development and Energy (StMWi) under the Bavarian Collaborative Research Programs (BayVFP), funding line Digitalization / Information and Communication Technology, by the German Federal Ministry for Economic Affairs and Climate Action of Germany (BMWK), and in part by the German Federal Ministry for Research, Technology, and Space (BMFTR).

We thank Anna Lanz and Tooba Islam for leading the annotation process, and Christopher Zak for creating the paper figures and visual assets.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report, 2023.

Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K., et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025.

Bansal, S., Arora, C., and Jawahar, C. V. My view is the best view: Procedure learning from egocentric videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

Bonnetto, A., Qi, H., Leong, F., Tashkovska, M., Rad, M., Shokur, S., Hummel, F., Micera, S., Pollefeys, M., and Mathis, A. Epfl-smart-kitchen: An ego-exo multi-modal dataset for challenging action and motion understanding in video-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Chen, D., Cahyawijaya, S., Ishii, E., Chan, H. S., Bang, Y., and Fung, P. What makes for good image captions? In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 1420–1437, November 2025. doi: 10.18653/v1/2025.findings-emnlp.75.

Chen, D., Kasarla, T., Bang, Y., Shukor, M., Chung, W., Yu, J., Bolourchi, A., Moutakanni, T., and Fung, P. Action100m: A large-scale video action dataset, 2026.

Cho, J. H., Madotto, A., Mavroudi, E., Afouras, T., Nagarajan, T., Maaz, M., Song, Y., Ma, T., Hu, S., Jain, S., et al. Perceptionlm: Open-access data and models for detailed visual understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Damen, D., Doughty, H., Farinella, G. M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., and Wray, M. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Darkhalil, A., Shan, D., Zhu, B., Ma, J., Kar, A., Higgins, R., Fidler, S., Fouhey, D., and Damen, D. Epic-kitchens visor benchmark: Video segmentations and object relations. *Advances in Neural Information Processing Systems*, 35: 13745–13758, 2022.

Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., and Florence, P. PaLM-E: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Feichtenhofer, C., Fan, H., Malik, J., and He, K. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., Chen, P., Li, Y., Lin, S., Zhao, S., Li, K., Xu, T., Zheng, X., Chen, E., Shan, C., He, R., and Sun, X. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N. A., Ma, W.-C., and Krishna, R. BLINK: Multimodal large language models can see but not perceive. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Gao, J., Sarkar, B., Xia, F., Xiao, T., Wu, J., Ichter, B., Majumdar, A., and Sadigh, D. Physically grounded vision-language models for robotic manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 12462–12469. IEEE, 2024.

Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400, 2024.

Grunde-McLaughlin, M., Krishna, R., and Agrawala, M. AGQA: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Guo, Q., De Mello, S., Bai, H., Hu, Y.-T., Jiang, Y., Akula, A., Tulyakov, S., Kautz, J., Wang, X. E., and Molchanov, P. RegionGPT: Towards region understanding vision language model. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Intelligence, P., Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., et al. $\pi_0.5$: a vision-language-action model with open-world generalization, 2025.

Ji, J., Krishna, R., Fei-Fei, L., and Niebles, J. C. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Khemlani, S., Tran, T., Gyory, N., Harrison, A. M., Lawson, W. E., Thielstrom, R., Thompson, H., Singh, T., and Trafton, J. G. Vision language models are unreliable at trivial spatial cognition, 2025.

Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., Vuong, Q., Kollar, T., Burchfiel, B., Tedrake, R., Sadigh, D., Levine, S., Liang, P., and Finn, C. OpenVLA: An open-source vision-language-action model. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2024.

Knab, P., Marton, S., Schubert, P. J., Guggiana, D., and Bartelt, C. Concepts in motion: Temporal bottlenecks for interpretable video classification, 2025.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, 2023.

Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., Wang, L., and Qiao, Y. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22195–22206, 2024.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., and Zhang, L. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., and Yi, L. HOI4D: A 4D egocentric

dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Mangalam, K., Akshulakov, R., and Malik, J. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023. Datasets and Benchmarks Track.

Meta FAIR. SAM 3: Segment anything with concepts, 2025.

Perrett, T., Darkhalil, A., Sinha, S., Emara, O., Pollard, S., Parida, K. K., Liu, K., Gatti, P., Bansal, S., Flanagan, K., Chalk, J., Zhu, Z., Guerrier, R., Abdelazim, F., Zhu, B., Moltisanti, D., Wray, M., Doughty, H., and Damen, D. Hd-epic: A highly-detailed egocentric video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23901–23913, June 2025.

Prasse, K., Jung, S., Bravo, I., Walter, S., Knab, P., Bartelt, C., and Keuper, M. From codebooks to vlms: Evaluating automated visual discourse analysis for climate change on social media, 2026.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., and Feichtenhofer, C. SAM 2: Segment anything in images and videos. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Salehi, M., Park, J. S., Yadav, T., Kusupati, A., Krishna, R., Choi, Y., Hajishirzi, H., and Farhadi, A. Actionatlas: A videoqa benchmark for domain-specialized action recognition. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. doi: 10.52202/079017-4364. Datasets and Benchmarks Track.

Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., and Yao, A. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Su, Y., Wang, Y., and Chau, L.-P. CaRe-Ego: Contact-aware relationship modeling for egocentric interactive hand-object segmentation. *Expert Systems with Applications*, 2026.

Tang, Y., Ding, D., Rao, Y., Zheng, Y., Zhang, D., Zhao, L., Lu, J., and Zhou, J. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1207–1216, 2019.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. Gemini: a family of highly capable multimodal models, 2023.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., et al. Gemma: Open models based on gemini research and technology, 2024.

Tu, C., Zhang, L., Chen, P., Ye, P., Zeng, X., Cheng, W., YU, G., and Chen, T. Favor-bench: A comprehensive benchmark for fine-grained video motion understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Wang, X., Kwon, T., Rad, M., Pan, B., Chakraborty, I., Andrist, S., Bohus, D., Feniello, A., Tekin, B., Frujeri, F. V., Joshi, N., and Pollefeys, M. HoloAssist: An egocentric human interaction dataset for interactive AI assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Wu, H., Li, D., Chen, B., and Li, J. Longvideobench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. doi: 10.52202/079017-0907. Datasets and Benchmarks Track.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report, 2025a.

Yang, J., Peng, W., Li, X., Guo, Z., Chen, L., Li, B., Ma, Z., Zhou, K., Zhang, W., Loy, C. C., and Liu, Z. Panoptic video scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18675–18685, June 2023a.

Yang, J., Zhang, H., Li, F., Zou, X., Li, C., and Gao, J. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V, 2023b.

Yang, J., Liu, S., Guo, H., Dong, Y., Zhang, X., Zhang, S., Wang, P., Zhou, Z., Xie, B., Wang, Z., Ouyang, B., Lin, Z., Cominelli, M., Cai, Z., Zhang, Y., Zhang, P., Hong, F., Widmer, J., Gringoli, F., Yang, L., Li, B., and Liu, Z. EgoLife: Towards egocentric life assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S.-F., and Yang, Y. Ferret: Refer and ground anything anywhere at any granularity. In *International Conference on Learning Representations (ICLR)*, 2024.

Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., and Tao, D. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 9127–9134, 2019.

Zhang, L., Zhou, S., Stent, S., and Shi, J. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

Zhou, L., Xu, C., and Corso, J. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, 2018.

A. Benchmark implementation details

A.1. Common conventions

Evaluation pipeline. Every task is evaluated on a deterministic, frozen set of examples derived from the released annotations. For temporal VQA, reference question-answer pairs are generated once offline using Gemini 2.5 Flash Lite and cached so that downstream evaluation remains reproducible. Each model is queried on the same frozen examples and scored with task-specific metrics. For open-ended tasks, LLM-based judging is applied as a separate post-hoc step, allowing predictions to be re-scored under different judge configurations without repeating inference.

Bounding-box representation. Bounding boxes are stored canonically as absolute (x_1, y_1, x_2, y_2) pixel coordinates in the frozen evaluation set. At prompt construction time, coordinates are rescaled to the integer range $[0, 1000]$ and reordered into the format expected by each provider (e.g. y_1, x_1, y_2, x_2 for Gemini, Gemma and GPT, x_1, y_1, x_2, y_2 for Qwen). This lets a single frozen dataset be reused across providers; model responses are always rescaled back to pixel coordinates before scoring.

Decoding. All models are queried at temperature 0.7, except GPT-5.4 mini, for which OpenAI does not expose custom sampling parameters for reasoning models. Reference-asset generation (temporal VQA question-answer pairs) also uses temperature 0.7; outputs are cached so that downstream evaluation remains deterministic.

Judge model. All G-Eval evaluations use Gemini 2.5 Flash as the judge via the DeepEval framework (Liu et al., 2023). Each criterion is scored on a 0–10 rubric and linearly rescaled to $[0, 1]$ for reporting.

A.2. Object-focused tasks

A.2.1. Phrase grounding

Examples. Candidate frames are restricted to those (a) that fall inside an annotated activity segment and (b) that carry at least one annotated relation. The surviving frames are evenly subsampled with a sampling fraction of 0.1, from which $1{,}000$ examples are shuffled and frozen with a fixed seed. For each frame, phrases are generated per category by a greedy selector that scores candidate descriptions by richness, preferring category together with color attributes together with a spatial relation over bare category names. Relations derived from the annotations are expressed through templates such as "*the* ⟨*color*⟩ ⟨*category*⟩ ⟨*relation*⟩ *the* ⟨*target*⟩". At most one phrase per category is retained per frame; objects with bounding-box area below 100 pixels are excluded.

System prompt.

```
You are an expert at analyzing video frames to locate specific objects.
For each object matching the description, output its bounding box.
Bounding box format: {coord_desc}, integers 0--{scale}. Order: {coord_order}
The final output should be a JSON object:
{"bboxes": [[a, b, c, d], [a, b, c, d]]}
Each inner list is exactly 4 integers. If no matching objects, return: {"bboxes":
[]}
```

User prompt. The user prompt attaches the frame and the generated phrase:

```
Locate all instances of: "{phrase}"
```

Metrics. We report mAP and AR averaged over IoU thresholds $[0.5{:}0.05{:}0.95]$, mAP@0.5, mAP@0.75, and size-stratified mAP/AR using the COCO size bins (Lin et al., 2014): small ($< 32^2$ px), medium (32^2–96^2 px), and large ($> 96^2$ px).

Full results. Table 8 reports the complete set of COCO-style detection metrics stratified by object size. Performance drops substantially on medium-sized objects across all models, confirming that object scale remains a key bottleneck for phrase grounding in BARISTA.

Table 8. **Phrase grounding (full results)** on 1000 frozen examples. mAP/AR over IoU $\in [0.5, 0.95]$ and at IoU 0.5, 0.75; size bins follow COCO conventions.

	All @[.5:.95]		Small		Medium		Large		mAP@50	mAP@75
Model	mAP	mAR	mAP	mAR	mAP	mAR	mAP	mAR	All	All
SAM 3	.362	.473	.084	.103	.192	.245	.366	.485	.393	.377
Gemini 3 Flash	$\mathbf{.534} \pm .008$	$\mathbf{.646} \pm .002$	$.012 \pm .001$	$.020 \pm .004$	$\mathbf{.237} \pm .019$	$\mathbf{.292} \pm .028$	$\mathbf{.567} \pm .011$	$\mathbf{.691} \pm .006$	$\mathbf{.697} \pm .017$	$\mathbf{.564} \pm .007$
GPT-5.4 mini	$.068 \pm .014$	$.114 \pm .017$	$.000 \pm .000$	$.000 \pm .000$	$.001 \pm .000$	$.003 \pm .002$	$.073 \pm .015$	$.124 \pm .018$	$.163 \pm .022$	$.048 \pm .011$
Gemma 4-31B	$.432 \pm .004$	$.555 \pm .005$	$.002 \pm .000$	$.005 \pm .001$	$.109 \pm .015$	$.147 \pm .019$	$.469 \pm .002$	$.605 \pm .003$	$.634 \pm .005$	$.473 \pm .010$
Qwen3.5-27B	$.518 \pm .005$	$.624 \pm .005$	$\mathbf{.021} \pm .021$	$\mathbf{.026} \pm .022$	$.193 \pm .036$	$.238 \pm .030$	$.549 \pm .007$	$.664 \pm .003$	$.680 \pm .016$	$.552 \pm .022$
Qwen3.5-9B	$.350 \pm .020$	$.454 \pm .017$	$.001 \pm .001$	$.002 \pm .000$	$.040 \pm .017$	$.056 \pm .015$	$.393 \pm .021$	$.507 \pm .018$	$.559 \pm .027$	$.374 \pm .037$

A.2.2. EGOCENTRIC HAND-OBJECT INTERACTION

Examples. Interactions are extracted from the annotated *human_actions* relations. The hand type (left, right, or both) is inferred from the source category; if the same target object appears simultaneously as the target of both a left-hand and a right-hand relation in the same frame, the two interactions are merged into a single *both*-hand interaction. The frozen evaluation set contains 1,000 frames, each with at least one interaction.

System prompt. The system prompt includes three inline exemplars demonstrating the right-hand, both-hands, and two-separate-interactions cases.

```
You are analyzing egocentric (first-person) video frames.  Identify every
hand-object interaction:  one or both of the camera wearer's hands actively
manipulating, holding, or touching an object.
For each interaction output:
- "hand_boxes":  a list of 1 or 2 bounding boxes --- one per interacting hand.
- "object_box":  a single bounding box for the object being interacted with.
- "hand_type":  which hand(s) are involved --- "left", "right", or "both".
Each interaction is a separate element.  If the same hand touches multiple objects,
report each interaction separately.  Bounding box format:  {coord_desc}, integers
0--{scale}.
Return ONLY a JSON array --- one element per interaction.  If none, return [].
```

User prompt. The user prompt attaches the frame without additional text.

Metrics. Predicted and ground-truth interactions are matched greedily one-to-one by object-box IoU at a threshold of 0.5, with highest-overlap predictions matched first. On matched pairs, hand boxes are themselves matched by a secondary greedy IoU assignment. We report interaction-level precision and recall, the mean IoU of matched hand and object boxes, the accuracy of the hand-type label on matched pairs, and a *no-detection rate* computed as the fraction of frames that contain at least one ground-truth interaction but for which the model emits no prediction. CaRe-Ego is evaluated on the same frozen evaluation set using its released inference pipeline.

Full results. Table 9 reports the complete set of HOI metrics, including hand-type classification accuracy.

A.2.3. REFERRING / REGION DESCRIPTION

Examples. References are assembled deterministically from the verified annotations. For each selected object, the reference includes: (1) the object category; (2) its attributes as key-value pairs (color, type, state); (3) its direct relations (e.g. "coffee cup under coffee machine", "right hand holds coffee cup"); and (4) a context block for each related object, including that object's own attributes and relations. Candidate frames are restricted to those inside annotated activity segments and subsampled with a stride of 60 frames. For each candidate frame, up to 10 objects are selected from those that possess both attributes and at least one spatial or functional relation; at most 50 examples per object category are retained. The frozen evaluation set contains 1,000 examples, each consisting of a single frame paired with a bounding box.

Table 9. **Egocentric HOI (full results)** on 1000 frozen frames. Matching: greedy by object-box IoU > 0.5. No-det. = No-detection rate (↓).

Model	Int. Prec.	Int. Rec.	Hand IoU	Obj. IoU	Hand-type Acc.	No-det. Rate
CaRe-Ego	.444	.547	.703	.839	.807	.071
Gemini 3 Flash	$.791 \pm .003$	$.711 \pm .011$	$\mathbf{.744} \pm .007$	$\mathbf{.892} \pm .001$	$.901 \pm .005$	$.028 \pm .007$
GPT-5.4 mini	$.203 \pm .015$	$.184 \pm .014$	$.483 \pm .007$	$.644 \pm .008$	$.833 \pm .006$	$.006 \pm .001$
Gemma 4-31B	$\mathbf{.799} \pm .003$	$.721 \pm .002$	$.707 \pm .001$	$.857 \pm .001$	$.837 \pm .003$	$.015 \pm .002$
Qwen3.5-27B	$.779 \pm .008$	$\mathbf{.729} \pm .006$	$.706 \pm .004$	$.864 \pm .002$	$\mathbf{.917} \pm .005$	$\mathbf{.004} \pm .001$
Qwen3.5-9B	$.715 \pm .006$	$.583 \pm .009$	$.640 \pm .003$	$.787 \pm .005$	$.905 \pm .003$	$.121 \pm .008$

System prompt.

```
You describe objects in video frames of an egocentric coffee-making scenario.  You
will be given a bounding box in {coord_desc} format, with integer coordinates
normalized to 0--{scale}.
Write one natural sentence that:
1.  Names the object using the most specific category you can determine.
2.  Includes key visual attributes you can observe.
3.  States the most important spatial or functional relation to another object.
Avoid incidental proximity (near, next to, beside).
4.  Does not mention background, surfaces, walls, or scenery.
Start your final response with ``Description:''  followed by your sentence.
```

User prompt. The user prompt attaches the frame and the query: "Describe the object inside bounding box [y1, x1, y2, x2]".

G-Eval rubric. Referring outputs are scored on a single *Correctness* criterion (0–10 scale, rescaled to $[0, 1]$). The evaluation prompt sent to the judge includes the domain vocabulary listing all canonical category names with short descriptions, enabling the judge to assess category specificity. The full criteria text provided to the judge is:

Evaluate whether the predicted referring expression correctly and specifically identifies the object in an egocentric coffee-making video frame. The reference is a structured annotation in the format:
Object: ⟨category⟩
Attributes: ⟨key=value, . . . ⟩
Relations: ⟨direct relations of the object⟩
Related objects: ⟨attributes and relations of related objects—supplementary context⟩

The "Relations" line lists all direct relations, but the prediction is a single sentence that is not expected to mention every one. The "Related objects" section is supplementary context for disambiguation.

A good prediction:
(1) Names the object with the correct or acceptably synonymous category—accept reasonable synonyms (e.g. "espresso machine" for "portafilter machine"), but penalize significantly less specific names (e.g. "machine" for "capsule machine").
(2) States key visual attributes of the main object correctly (color, state).
(3) Includes at least one primary spatial or functional relation to another object (e.g. "held by a hand", "under a capsule machine", "attached to a portafilter machine").
(4) Is appropriately concise—a single focused sentence without scene description, background, bounding-box coordinates, or brand names.

Extra correct details, equivalent wording, and minor rephrasing should never reduce the score.

Domain vocabulary (canonical category names; accept reasonable synonyms, but prefer more specific names over generic ones): [*full list of category names with descriptions omitted for brevity*]

Correctness.
Score 0–3: Prediction names a completely different object category that contradicts the reference, or is unintelligible/empty.
Score 4–7: Correct object but with notable errors: overly generic category (e.g. "coffee machine" instead of "capsule machine"),

wrong attribute, wrong relation, omits all relations, or excessively verbose.
Score 8–10: Correctly identifies the object with its key attributes and at least one correct primary spatial/functional relation in a concise single sentence. Equivalent wording and extra correct details are acceptable.

A.3. Semantic tasks

A.3.1. Activity recognition

Examples. Each example consists of $F{=}4$ frames sampled uniformly from a single annotated activity segment. Segments shorter than F frames are discarded; at most 20 examples per activity class are retained to enforce category balance. 500 examples are sampled across videos with a fixed seed.

Distractor construction. Each question presents $K{=}25$ choices (one correct, 24 distractors). When available, at least 4 distractors are guaranteed to share the same verb as the correct answer (e.g. *close bottle*, *close cabinet*, *close coffee machine*), forcing the model to discriminate based on the manipulated object rather than the action alone. The remaining distractors are drawn randomly from the full activity vocabulary; all 25 choices are then shuffled with a deterministic per-segment seed.

System prompt.

```
You are an expert at fine-grained activity recognition from video frames.  Focus
on hand-object interactions and visible state changes across consecutive frames
(e.g., direction of motion, whether an object is being placed or removed).  You may
provide a brief rationale, but the final line must be exactly ANSWER: <CHOICE_ID>.
```

User prompt. The user prompt attaches the four frames in order and poses the question with the 25 activity labels as choices A through Y, each rendered from the display name of the corresponding activity class:

```
The 4 frames below are sampled uniformly from an activity segment (Frame 1
earliest, Frame 4 latest).  Which activity label best describes the segment?
Choices:
A. ...
B. ...
...
```

Metrics. Responses are parsed by extracting the choice identifier from an "ANSWER: ¡ID¿" line (case-insensitive, allowing optional rationale before the final line). We report top-1 accuracy over all examples (parse failures scored as incorrect).

A.3.2. Relation extraction

Examples. Candidate frames are subsampled with a temporal stride of 300 frames and filtered to those carrying at least one annotated relation. 1,000 frames are shuffled and frozen with a fixed seed.

Set-of-mark overlay. Each frame is shown twice: once as the original RGB frame and once as a *set-of-mark* overlay in which the instance masks of annotated objects are blended with the frame at alpha 0.5, their boundaries are drawn at alpha 0.9, and each instance is labeled with an integer identifier at font size 20. A legend mapping identifier to category is included in the user prompt. The allowed relation vocabulary is derived from all video annotations and appended to the system prompt, listing for each relation type the valid set of value strings.

System prompt.

```
You extract pairwise relations between objects in a coffee-making scene.  Each
object has a numeric ID on the set-of-mark image and in the ID-to-category list.
The final response must be a JSON array of relation objects.  If there are no
relations, return [].  Each element must be an object with keys:  source_id
(integer), target_id (integer), relation_type (string), value (string).  Use only
relation_type and value strings that appear together under the same relation_type
```

```
in the user message.
Example: [{"source_id": 1, "target_id": 2, "relation_type": "position", "value":
"part of"}]
```

User prompt. The user prompt attaches the original frame, the set-of-mark overlay, and the identifier-to-category mapping.

Metrics. Predicted and ground-truth relations are compared as multisets of tuples (s, t, r, v) after case- and whitespace-normalization of all string fields. We report micro-averaged precision, recall, and F1 pooled over all frames, together with the same metrics broken down by relation type.

Full results. Table 10 reports the complete precision, recall, and F1 breakdown. Qwen3.5-27B obtains the highest recall (.799), while Gemma 4-31B achieves the highest precision (.736).

Table 10. **Relation extraction (full results)**. Micro P/R/F1 under exact multiset matching; per-type micro F1.

	Overall			Per-type micro F1	
Model	P_{micro}	R_{micro}	F1_{micro}	human_actions	position
Gemini 3 Flash	$.655 \pm .006$	$.697 \pm .009$	$.675 \pm .008$	$\mathbf{.675} \pm .004$	$.710 \pm .007$
GPT-5.4 mini	$.546 \pm .009$	$.586 \pm .008$	$.565 \pm .008$	$.534 \pm .011$	$.569 \pm .008$
Gemma 4-31B	$\mathbf{.736} \pm .002$	$.767 \pm .002$	$\mathbf{.751} \pm .002$	$.673 \pm .012$	$\mathbf{.758} \pm .001$
Qwen3.5-27B	$.699 \pm .006$	$\mathbf{.799} \pm .003$	$.745 \pm .005$	$.626 \pm .006$	$.756 \pm .004$
Qwen3.5-9B	$.587 \pm .005$	$.642 \pm .002$	$.613 \pm .004$	$.573 \pm .010$	$.630 \pm .005$

A.3.3. Temporal visual question answering

Examples. Reference QA pairs are produced offline by Gemini 2.5 Flash at temperature 0.7. For every candidate clip the generator detects annotation diffs between the clip's start and end windows—which object attributes changed, which hand-contact sequences occurred, which spatial relations shifted, and which activity labels transitioned—and instantiates one QA pair per non-empty category. Clip starts are strided every 144 frame indices (non-overlapping); each clip contains 6 keyframes spaced 24 frame indices apart, spanning four seconds at 30 fps. The generator's system prompt is:

```
You are creating category-tagged question--answer pairs for a visual QA benchmark
about coffee-making videos captured from an egocentric viewpoint. [...] Generate
exactly 1 question--answer pair for the assigned category. The question should
naturally encompass all listed changes of that type. [...] Return ONLY a valid
JSON object: {"question": "...", "answer": "...", "category": "..."}
```

QA pairs are cached to disk and reused verbatim across every subsequently evaluated model. 300 pairs are sampled uniformly across the four categories (*state change*, *action transition*, *relation change*, *activity progression*; up to 100 per category) to form the frozen evaluation set.

System prompt.

```
You are an expert at visual question answering for coffee-making videos. You
will receive a sequence of ordered frames sampled from a video clip, along with
a question about the content. Analyze the frames carefully and provide a detailed,
accurate answer. Your answer should be specific, clear, and based solely on what
is visible in the frames. Start your final response with "Answer:" followed by
your answer.
```

User prompt. The user prompt attaches the frames (up to six) in order and poses the reference question verbatim. Answers are extracted by taking the text after the final "Answer:" marker (case-insensitive).

G-Eval rubrics. Both temporal VQA criteria are scored on the 0–10 scale defined below and are rescaled to $[0, 1]$ for reporting.

Accuracy. Are all claims in the prediction factually correct per the reference? Only penalize claims that are directly contradicted—wrong change direction, wrong object name, or wrong contact state. Do not penalize omissions, extra details, or equivalent wording. Contact-action synonyms (touching/holding/gripping the same object) are the same state. Machine synonyms (coffee machine / espresso machine / capsule machine) are always acceptable. For action-transition questions, evaluate whether the contact targets in the prediction are correct—do not penalize for omitting or including temporal anchors like "at the start" or "by the end", since the hand may appear at any point in the clip.

Score 0–3: The prediction contains clearly wrong claims that directly contradict the reference—reversed change directions, wrong object names, or a stated contact state that is inconsistent with the reference.
Score 4–7: The prediction is mostly correct but includes at least one claim that directly contradicts the reference (e.g. wrong change direction or wrong object name). Additional details not in the reference should not place the answer here.
Score 8–10: All key claims in the prediction are correct and consistent with the reference. Additional correct details, plausible intermediate steps, or non-contact observations not in the reference are acceptable. Equivalent terminology is acceptable.

Completeness. Does the prediction cover every change listed in the reference? Do not penalize incorrect claims—only missing ones. Idle hand states between contact events do not need to be stated explicitly. Per-hand attribution (left vs. right) is not required. Semantically equivalent descriptions are acceptable. For action-transition questions, evaluate whether every contact target in the reference is mentioned in the prediction—ignore temporal anchors like "at the start" or "by the end" since the hand may appear at any point in the clip.

Score 0–3: The prediction omits most changes—it gives a vague or generic response that skips the specific objects, states, or transitions from the reference.
Score 4–7: The prediction mentions some changes but misses at least one important object, state transition, or activity step present in the reference.
Score 8–10: The prediction covers all changes from the reference. Minor rephrasing is acceptable.

Full results. Table 11 breaks down temporal VQA performance by question category. GPT-5.4 mini performs particularly well on state-change questions, while relation-change questions are easiest across models and action transitions are hardest.

Table 11. **Temporal VQA per-category mean** (G-Eval mean of Accuracy and Completeness, in $[0, 1]$).

Model	Relation change	Action transition	State change	Activity progression
Gemini 3 Flash	$0.786 \pm .019$	$0.548 \pm .012$	$0.657 \pm .010$	$0.603 \pm .023$
GPT-5.4 mini	$0.760 \pm .023$	$0.525 \pm .004$	$\mathbf{0.817} \pm .015$	$0.543 \pm .013$
Gemma 4-31B	$0.785 \pm .007$	$0.470 \pm .026$	$0.630 \pm .015$	$\mathbf{0.607} \pm .023$
Qwen3.5-27B	$\mathbf{0.826} \pm .010$	$\mathbf{0.575} \pm .034$	$0.743 \pm .009$	$0.585 \pm .008$
Qwen3.5-9B	$0.793 \pm .043$	$0.486 \pm .019$	$0.732 \pm .033$	$0.509 \pm .019$

A.4. Dataset and Code Release

We provide both the dataset and the accompanying code for reproducibility. The dataset is available at `https://huggingface.co/datasets/ramblr/BARISTA`. It contains the BARISTA data, including videos, annotations, and splits. The code is available at `https://github.com/Ramblr-GmbH/BARISTA`. It contains the scripts required to reproduce the benchmark construction and evaluation protocol, including task sampling, prompt generation, model outputs, and metric computation.

A.5. Runtime and resources

All experiments with open-weight models were run on the same serving setup within each model scale. For Qwen3.5-27B and Gemma 4-31B, each run used two NVIDIA H200 GPUs with 128 GB system RAM in total. We served one VLM instance per GPU and used data parallelism (degree 2); requests were processed with a total concurrency of 4. For Qwen3.5-9B, we used the same serving configuration, but ran the model on NVIDIA RTX PRO 6000 GPUs, as the smaller model required less memory.

Token usage. We report the average number of input and output tokens per example for each model–task pair in Table 12. Token counts reflect provider-specific tokenization and include any system-prompt overhead.

Table 12. **Token usage overview.** Average token usage per example across all benchmark tasks and models. In = input tokens, Out = output tokens.

	Gemini 3 Flash		GPT-5.4 mini		Qwen3.5-27B		Gemma 4-31B		Qwen3.5-9B	
Task	In	Out	In	Out	In	Out	In	Out	In	Out
Grounding	1236	228	1217	173	2365	704	1272	425	2365	894
Hand-object	1581	104	1494	174	2626	1543	1628	602	2626	2026
Referring	1254	385	1214	165	2424	1443	1288	433	1615	1148
Relation extraction	2521	524	2547	332	4522	3829	2573	1134	4526	4911
Activity recognition	4617	559	4604	185	9267	1911	4661	945	9267	2473
Temporal VQA	6585	304	6809	136	12034	1792	6643	545	12058	2254

A.6. BARISTA's characteristics

Figure 3 presents key distributional properties of BARISTA. Capsule machines account for the majority of recordings, followed by fully automatic and portafilter machines (a). Annotated frames typically contain 6–10 simultaneously tracked objects (b), reflecting the dense, multi-object nature of coffee preparation scenes. Most videos fall in the 1–3 min range (c), while activity segment durations range from sub-second manipulations to under one minute (d).

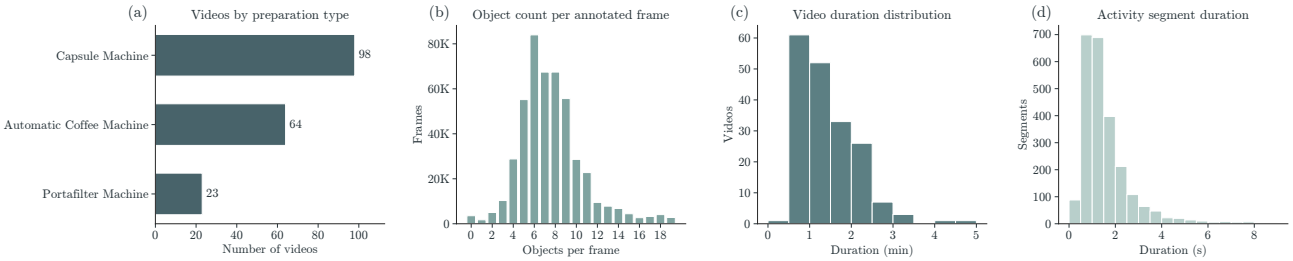

Figure 3. **Dataset overview.** (a) Video distribution across preparation types. (b) Number of annotated objects per frame. (c) Video duration distribution. (d) Activity segment duration.

Figure 4 shows the annotation distributions. Object category counts (a) exhibit a long tail: high-frequency categories such as *button*, *cup*, and *coffee machine* appear in nearly every video, while specialized items occur only in specific preparation styles. Activity classes (b) are similarly skewed: common actions such as *Place cup* and *Press button* occur hundreds of times. Relation predicates (c) are grouped by type: hand-object interactions (*holds*, *touches*) dominate the *human actions* group, while spatial predicates (*part of*, *under*, *contains*) capture the static scene structure.

A.7. LLM as a judge: Cross-judge robustness

To assess the sensitivity of G-Eval scores to judge identity, we re-score all five models on both judge-dependent tasks—referring expression generation (1,000 examples) and temporal VQA (300 examples)—using GPT-5.4 mini as an alternative judge, replacing Gemini 2.5 Flash used in the main evaluation. The two judges belong to different model families (Google vs. OpenAI), providing a cross-family consistency check.

Ranking preservation. Table 13 and Table 14 report per-model scores under both judges. We compute Kendall τ-b rank correlation across the five models for each metric. On referring (Correctness), the model ranking is identical under both judges ($\tau = 1.0$). On temporal VQA, the ranking is preserved for mean score ($\tau = 1.0$) and Completeness ($\tau = 1.0$); for Accuracy alone, a single adjacent swap occurs between Gemini 3 Flash and GPT-5.4 mini—models that were already within 0.004 of each other under the primary judge—yielding $\tau = 0.8$.

Absolute calibration shift. The GPT-5.4 mini judge assigns systematically lower scores than Gemini 2.5 Flash (mean $\Delta \approx -0.07$ on referring, -0.08 on VQA), indicating a stricter calibration. However, the shift is approximately uniform across models: the standard deviation of Δ across models is 0.023 on referring and 0.002 on VQA mean, indicating that the offset is judge-specific rather than model-specific.

Table 13. **Referring expression** G-Eval Correctness under two judges. Model ranking is identical ($\tau = 1.0$).

Model	Gemini 2.5 Flash	GPT-5.4 mini	Δ
Gemini 3 Flash	.757	.643	$-.114$
Gemma 4-31B	.669	.598	$-.071$
Qwen3.5-27B	.623	.542	$-.081$
Qwen3.5-9B	.569	.519	$-.050$
GPT-5.4 mini	.302	.320	$+.018$

Self-preference probe. Gemini 3 Flash—evaluated by its own family's judge in the primary setup—retains rank 1 on referring under the cross-family GPT-5.4 mini judge (0.643 vs. next-best 0.598). Conversely, GPT-5.4 mini's own referring predictions do not receive a score boost when judged by GPT-5.4 mini (0.320 vs. 0.302 under Gemini), remaining last-ranked under both judges. These observations provide evidence against systematic self-preference effects for either judge family on the tasks and rubrics used in BARISTA.

Per-category VQA analysis. Table 15 reports the per-category mean scores under both judges. The cross-judge ranking is consistent within categories: relation-change questions are easiest under both judges, and action transitions remain hardest.

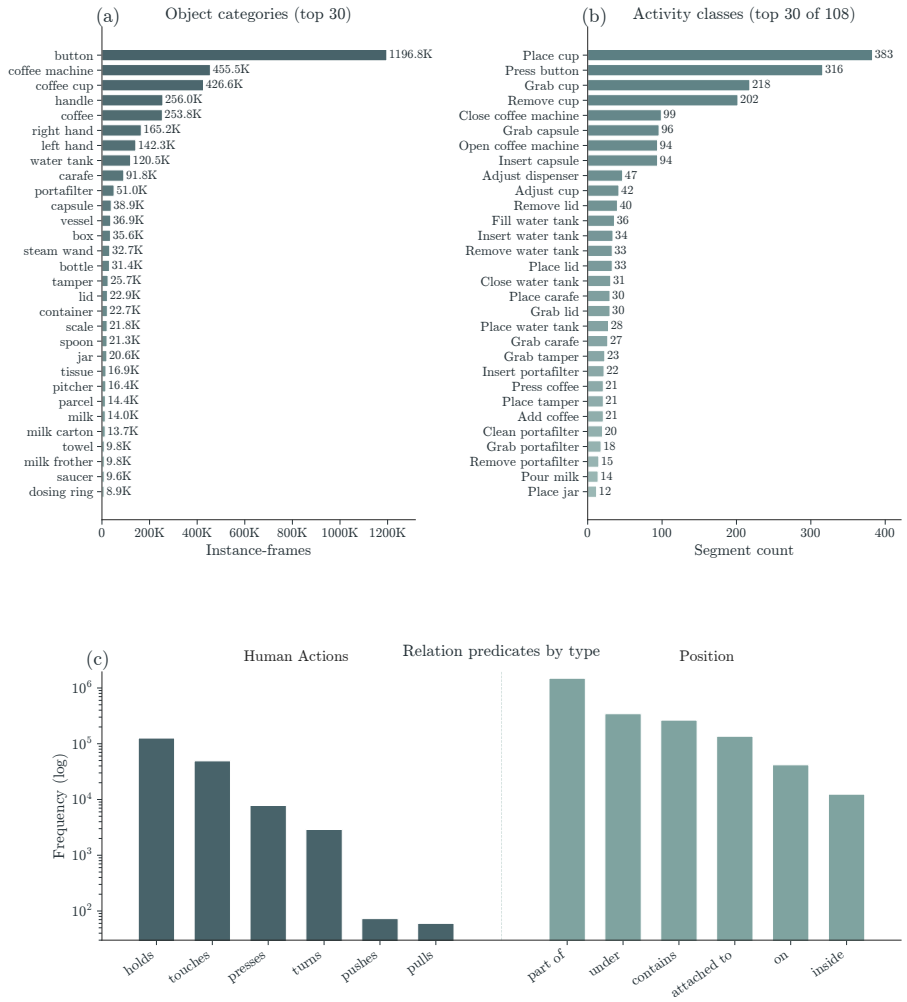

Figure 4. **Annotation distributions.** (a) Object category frequency (top 30 of 46 categories). (b) Activity class frequency (top 40 of 108 classes). (c) Relation predicate frequencies by type (log scale).

Table 14. **Temporal VQA** G-Eval scores under two judges. Ranking is preserved for Mean ($\tau = 1.0$) and Completeness ($\tau = 1.0$); Accuracy has one adjacent swap ($\tau = 0.8$).

	Mean			Accuracy			Completeness		
Model	Gemini	GPT	Δ	Gemini	GPT	Δ	Gemini	GPT	Δ
Qwen3.5-27B	.702	.622	−.080	.727	.596	−.131	.683	.648	−.035
GPT-5.4 mini	.683	.607	−.076	.701	.586	−.114	.671	.627	−.044
Gemini 3 Flash	.674	.599	−.075	.705	.577	−.129	.650	.620	−.030
Qwen3.5-9B	.666	.587	−.079	.691	.569	−.121	.645	.604	−.041
Gemma 4-31B	.647	.570	−.076	.665	.552	−.113	.630	.588	−.041

Table 15. **Temporal VQA per-category mean** under both judges.

	Gemini 2.5 Flash judge				GPT-5.4 mini judge			
Model	Rel. chg	State chg	Act. prog	Act. trans	Rel. chg	State chg	Act. prog	Act. trans
Qwen3.5-27B	.823	.743	.584	.596	.786	.684	.510	.357
GPT-5.4 mini	.772	.804	.535	.522	.744	.735	.455	.327
Gemini 3 Flash	.805	.653	.623	.535	.781	.562	.547	.382
Qwen3.5-9B	.841	.714	.527	.470	.768	.629	.463	.349
Gemma 4-31B	.778	.638	.601	.468	.725	.544	.542	.336

Summary. Model rankings on both judge-dependent tasks are robust to judge identity ($\tau \geq 0.8$ across all metrics, $\tau = 1.0$ on three of four). The single observed swap involves two models separated by < 0.005 under the primary judge. These results support the validity of the G-Eval protocol used in the main evaluation and indicate that the reported rankings reflect genuine performance differences rather than judge-specific artifacts.

A.8. Cross-task decomposability analysis

A core design claim of BARISTA is that its task families isolate complementary capabilities, so that performance on one part of the benchmark need not track performance on another. We test this claim post-hoc on the benchmark predictions.

Setup. For every task we compute a per-example score (mean best IoU for grounding, per-frame interaction F_1 for HOI, tuple F_1 for relation extraction, $0/1$ correctness for activity MCQ, G-Eval mean for referring and temporal VQA), and pool these scores across examples and all 3 seeds to obtain a single mean per (model, task) pair. We then compute Kendall τ-b on the 5-model rankings induced by every pair of tasks. A high τ between two tasks means they order the evaluated models almost identically (and so likely measure the same capability); a low or negative τ means one task ranks models in a way the other cannot predict.

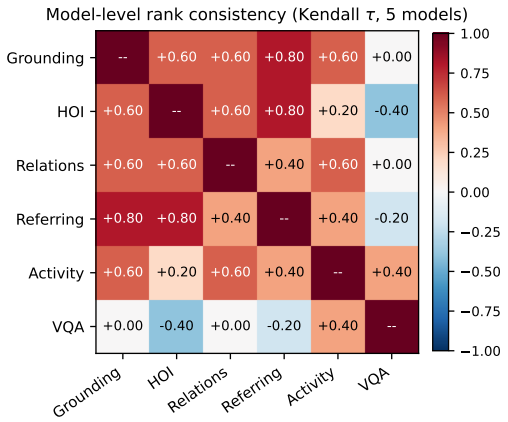

Figure 5. **Cross-task decomposability.** Kendall τ-b between model rankings across task pairs. The structural cluster is internally consistent, while Temporal VQA is decoupled from object-focused capabilities.

Result. Table 16 and Figure 5 report the full τ matrix. The four object-focused / structural tasks (grounding, HOI, relations, referring) form an internally consistent cluster ($\tau \geq 0.6$ on 5 of 6 pairs). Activity recognition is moderately aligned with the structural cluster ($\tau \in [0.2, 0.6]$). Temporal VQA, by contrast, is decoupled from grounding (τ=0.0), HOI (τ=−0.4), relations (τ=0.0) and referring (τ=−0.2); it correlates weakly only with activity recognition (τ=0.4). Although based on a small number of models, this rank analysis provides initial evidence that BARISTA does not collapse to a single "video competence" axis, but instead captures multiple distinct dimensions of video understanding. Concretely, GPT-5.4 mini ranks last on every object-focused task yet ranks among the top two on temporal VQA, while the strongest grounding models rank lower on VQA — a pattern that would be invisible to any single-axis benchmark.

Table 16. **Model-level rank consistency.** Kendall τ-b between the rankings of the five evaluated VLMs induced by each pair of tasks (computed on per-task pooled mean scores).

	Grounding	HOI	Relations	Referring	Activity	VQA
Grounding	–	+0.6	+0.6	+0.8	+0.6	+0.0
HOI	+0.6	–	+0.6	+0.8	+0.2	−0.4
Relations	+0.6	+0.6	–	+0.4	+0.6	+0.0
Referring	+0.8	+0.8	+0.4	–	+0.4	−0.2
Activity	+0.6	+0.2	+0.6	+0.4	–	+0.4
VQA	+0.0	−0.4	+0.0	−0.2	+0.4	–

A.9. Per-machine-type breakdown

BARISTA spans three preparation styles—capsule, portafilter, and fully automatic—that differ in workflow complexity, object set, and interaction patterns. We report per-task mean scores stratified by preparation type to verify whether these stylistic differences surface in model performance.

Results. Table 17 and Figure 6 report the full breakdown. Performance varies meaningfully across preparation types in a task-dependent manner. Portafilter videos yield the highest grounding scores (mean 0.636 vs. 0.507 for capsule), likely because the portafilter workflow involves larger, more distinctive objects at closer range. Conversely, fully automatic machines produce the best scores on relation extraction (0.727 vs. 0.619 capsule) and HOI (0.698 vs. 0.614 portafilter), reflecting a simpler, more constrained interaction graph. Activity recognition is easiest for portafilter (0.742) yet hardest for fully automatic (0.677), suggesting that models can recognize the distinctive manual steps (tamping, locking) more reliably than the subtler stages of an automated cycle. These divergent patterns confirm that the three preparation styles expose complementary performance characteristics rather than uniformly scaling difficulty.

Table 17. **Per-machine-type mean scores.** Average per-example score across all five models, stratified by preparation style and task (scale $[0, 1]$, higher is better). Scores are: mean best IoU (Grounding), interaction F_1 (HOI), tuple F_1 (Relations), G-Eval mean (Referring, VQA), accuracy (Activity).

Preparation	Grounding	HOI	Relations	Referring	Activity	VQA
Capsule	0.507	0.624	0.619	0.567	0.709	0.669
Portafilter	0.636	0.614	0.638	0.595	0.742	0.628
Fully Auto.	0.579	0.698	0.727	0.606	0.677	0.714

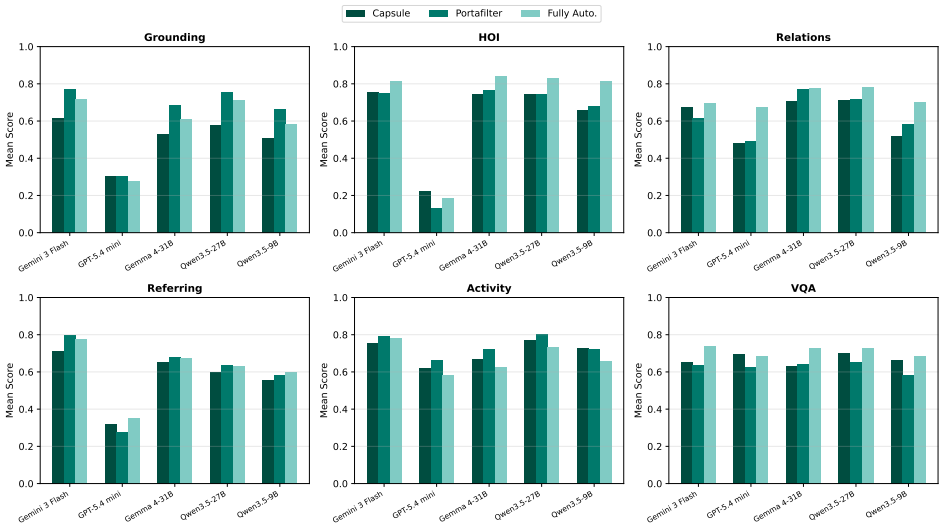

Figure 6. **Per-machine-type breakdown.** Mean per-example score by preparation style for each task and model. The three preparation styles expose different performance profiles: portafilter excels in grounding and activity recognition, while fully automatic dominates relation extraction and HOI.