

RetailAction: Dataset for Multi-View Spatio-Temporal Localization of Human-Object Interactions in Retail

Davide Mazzini Alberto Raimondi Bruno Abbate Daniel Fischetti David M Woollard
STANDARD.AI

{davide, alberto, bruno, dan, david.woollard}@standard.ai

Abstract

We introduce RetailAction, a novel dataset designed for multi-view spatio-temporal localization of human-object interactions in retail stores. Existing datasets either provide video-level action classification only (without spatio-temporal localization), or, when such annotations are present, they are limited in scale and not specific to the retail sector, often lacking real-world store data. RetailAction addresses these limitations by focusing on interactions between actual customers and store products, captured from multiple top-view cameras in 10 different real-world convenience stores. The dataset consists of 21,000 samples, each containing two synchronized videos with a total duration of 41 hours. In addition to the videos, the dataset includes annotations detailing precise interaction points for both views, temporal ranges, and action categories for each interaction. In this paper, we describe the data collection process and we provide an analysis of the dataset’s statistics. We also present a baseline model for spatio-temporal localization of interaction points and compare different state-of-the-art backbones. Finally, we present a novel set of evaluation metrics tailored to this use case. RetailAction aims to facilitate research on fine-grained action recognition and localization, offering a valuable resource for developing advanced retail analytics applications.

1. Introduction

Recent advances in computer vision have improved the understanding of human-object interactions, yet the retail domain remains underexplored. Fine-grained analysis of customer behavior is crucial for applications such as data analytics, product placement optimization, shopper behavior analysis, and autonomous checkout. However, existing datasets fail to adequately address the unique challenges of retail environments, where precisely localizing product-shopper interactions in both time and space is essential for all downstream applications.

Traditional large-scale action recognition datasets, such as Kinetics [3, 4, 13] and ActivityNet [6], primarily focus on video-level classification, lacking spatio-temporal localization. While AVA-Kinetics [16] includes sparse keyframe annotations, datasets like UCF101-24 [21] and JHMDB51-21 [12] provide denser spatio-temporal tubelets, though they remain limited in scale. Retail-specific datasets, such as MERL Shopping [20] and RetailVision [1], are constrained by scale, annotation granularity, and the absence of multi-view perspectives necessary for handling occlusions and cluttered settings.

To bridge this gap, we introduce **RetailAction**, a large-scale dataset designed for multi-view spatio-temporal localization of human-object interactions in real-world retail environments. Captured from operational convenience stores, it features synchronized multi-view video recordings of authentic customer interactions. The key contributions of this work include:

- 1. Novel Interaction-Centric Annotation Paradigm:** Unlike traditional spatio-temporal annotations that track the person performing an action, RetailAction introduces *interaction point annotations*—precisely marking where an object is handled. This shift enables a more fine-grained understanding of customer-product interactions, which is critical for retail analytics.
- 2. Large-Scale, Diverse, and Real-World Retail Data:** RetailAction consists of 21,000 annotated samples, totaling 41 hours of video, capturing unscripted customer behavior in 10 operational retail stores. The dataset features a diverse set of more than 10,000 unique shoppers, ensuring broad coverage of different shopping patterns and interaction styles. To collect and annotate these data at scale, we use a semi-automated pipeline that integrates 3D pose tracking, kinematic interaction detection, and dynamic frame rate adjustments, significantly improving annotation efficiency.
- 3. Multi-View Spatio-Temporal Annotations:** Each interaction is annotated across two synchronized camera views, providing detailed 2D interaction points and pre-

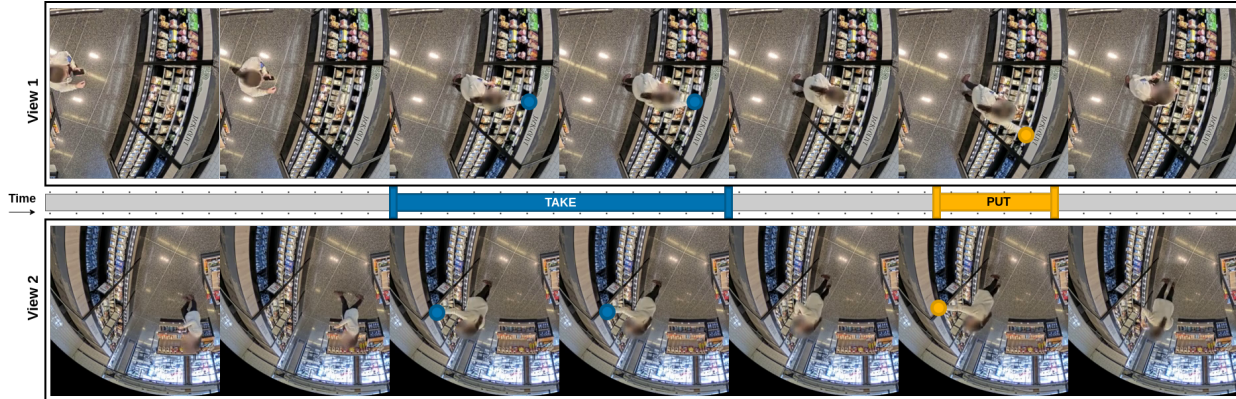


Figure 1. An annotated sample from the RetailAction dataset, consisting of two synchronized video streams from different cameras. The videos capture real people interacting in actual retail stores. The subject performs two actions: taking an item and placing another on a shelf. Each action is annotated with a categorical label (take, put, touch), a temporal range, and spatial coordinates in both views.

cise temporal boundaries. This facilitates robust action localization, even in crowded and occluded settings.

4. **Benchmarking and Baseline Models:** We developed a DETR-based model specifically tailored for the task of spatio-temporal localization of interactions from multiple cameras. In addition to this, we also developed a set of metrics specifically designed for the dataset and the task at hand. We evaluated our approach on the dataset alongside multiple state-of-the-art neural backbones.

RetailAction sets a new standard for retail-focused action localization, providing a foundation for advancing research in shopper behavior analysis and spatio-temporal interaction understanding.

2. Related Work

Existing datasets either focus on action classification — assigning predefined labels to entire clips — or support spatio-temporal localization by identifying when and where actions occur. Despite substantial progress, they often exhibit limitations for applications in retail settings.

Action Classification Datasets – In the realm of action classification, numerous datasets have been developed; however, most of these datasets fall short in terms of providing the spatial and temporal annotations that are essential for a retail-centric focus.

The *Kinetics* series [3, 4, 13] offers extensive video collections and is currently regarded as benchmark for action classification tasks. In addition to Kinetics, there are smaller datasets that encompass a wider variety of actions, such as *HMDB51* [15], *Moments in Time* [18], and *Something-Something-V2* [9]. Notably, the latter specifically focuses on human-object interactions from an egocentric viewpoint. Despite their contributions, none of these datasets are tailored to the retail context or provide the nec-

essary action localization information across both spatial and temporal dimensions.

Spatial and Temporal Localization Datasets – Some datasets, such as *THUMOS14* [11], *ActivityNet* [6], *HACS Segment* [24] and *FineAction* [17] simply add temporal annotations alongside classification, without supplying the spatial location of actions within the videos.

Other datasets also provide spatial information. Typically, this annotation is given in the form of bounding boxes around people, which are tracked over time. Some datasets, such as *AVA-Kinetics* [16], include only spatial labels, providing bounding boxes around people and specifying the action each person is performing, without indicating the exact moment the action occurs. Other datasets provide both spatial and temporal annotations. Among them, there are also large-scale examples: for instance, *AVA* [10] dataset includes sparse temporal annotations, consisting of individual annotated frames at separate time points, while the *UCF101-24* [21] and *JHMDB-21* [12] datasets provide continuous temporal annotations in the form of intervals. However, the focus in these datasets is always on the person rather than the point of interaction with objects, and the types of actions considered differ from those commonly found in retail scenarios, such as “take” and “put,” where knowing the exact location of the interaction is crucial.

Datasets for Retail – Retail-specific datasets remain scarce, particularly those with detailed spatio-temporal action annotations. The *MERL Shopping* [20] dataset captures retail interactions in a controlled laboratory setting but lacks temporal annotations. *RetailVision* [1] is a dataset that includes both spatial and temporal labels from shopping cart-mounted cameras, but its reliance on this viewpoint limits its generalizability to other retail settings such as ceiling-mounted cameras.

Dataset	Videos	Tot. Hours	Data source	Point of View	Temporal	Spatial	Retail	Multi-view
Kinetics-700 [4]	650k	1.8k	YouTube	TPV	-	-	-	-
UCF101 [21]	13,320	27	YouTube	TPV	-	-	-	-
SSV2 [9]	220k	246	Crowdsourced	Egocentric	-	-	-	-
HMDB51 [15]	6,849	2.5	YouTube	TPV	-	-	-	-
Moments in Time [18]	1M	833	Web sources	TPV	-	-	-	-
THUMOS14 [11]	413	0.49	YouTube	TPV	✓	-	-	-
ActivityNet [6]	20k	272	Web	TPV	✓	-	-	-
HACS Segment [24]	49k	449	YouTube	TPV	✓	-	-	-
FineAction [17]	17k	33	Other datasets + YT	TPV	✓	-	-	-
AVA [10]	403	100	YouTube	TPV	sparse	✓	-	-
AVA-Kinetics [16]	238k	761	Web	TPV	sparse	✓	-	-
UCF101-24 [21]	3207	6.5	YouTube	TPV	✓	✓	-	-
JHMDB51-21 [15]	928	0.3	YouTube	TPV	✓	✓	-	-
MERL Shopping [20]	96	3.2	Lab setting	Top view	✓	-	✓	-
RetailVision [1]	271	3.3	Lab setting	Shopping cart	✓	✓	✓	-
RetailAction - Ours	(2x) 21k	41	Real-world stores	Top view	✓	✓	✓	✓

Table 1. Comparison of action detection datasets in literature. Top section lists generic datasets, whereas bottom section highlights retail-focused datasets. *Sparse temporal label* indicates datasets where actions are annotated at keyframes only (e.g., 1 frame per second for AVA and 1 frame per video for Kinetics). *TPV* (Third-person view) represents handheld camera perspectives, *Top view* corresponds to ceiling-mounted cameras, and *Shopping cart* refers to an egocentric-like setup with cameras attached to shopping carts. Our RetailAction dataset uniquely captures real-world retail interactions with multi-view spatio-temporal annotations, enabling fine-grained analysis of human-object interactions in stores.

While numerous action recognition datasets exist, all of them provide only a single video per action. In contrast, our dataset includes two synchronized videos capturing the same action from different viewpoints, ensuring a more comprehensive perspective. Moreover, all our recordings are captured from a top-view perspective, which guarantees clear visibility of both the person and their interactions with objects. Additionally, our dataset is collected in real retail stores with actual customers rather than actors, making it representative of real-world shopping behavior. This design makes our dataset particularly complete and valuable for studying shopper interactions in retail environments.

3. Dataset collection

RetailAction dataset is collected in 10 real-world medium-to-small-sized convenience stores across various locations in the United States over the course of multiple years. The data correspond to actual customer visits. All shoppers were given notice of the recordings as well as signed terms of service with the company collecting the data. Moreover, the videos and their attached metadata have been anonymized to ensure that no individual could be identified. The recording devices used are 360-degree top-view cameras operating at 30 FPS with a resolution of 2880×2880. The cameras were mounted on the ceilings of the stores at an average

height of 2.5 meters and strategically positioned to ensure that every point within the area of interest is covered by at least two different cameras.

Given that each store is equipped with multiple cameras recording continuously throughout the day, the total volume of raw data is prohibitively large (often multiple TB per day). To generate a suitable dataset for action recognition, it is essential to develop an efficient method to manage this data, filtering out irrelevant footage while preserving meaningful interactions. Manually labeling all video streams to extract shopper-shelf interactions is impractical, especially when aiming to build a dataset as large as ours. Since actual shopper interactions with shelves represent only a fraction of the footage, and only a subset of cameras is needed to capture these interactions effectively, we implemented an advanced system capable of automatically detecting relevant time intervals and selecting the most suitable camera views. Additionally, our system is designed to operate in near real-time, further enhancing the efficiency of video storage and retrieval while handling the complexity of continuous data processing.

In this section, we provide a high-level overview of the data collection pipeline to convey the nature and complexity of our data collection approach and ensure transparency. Full reproducibility of the data collection pipeline lies beyond

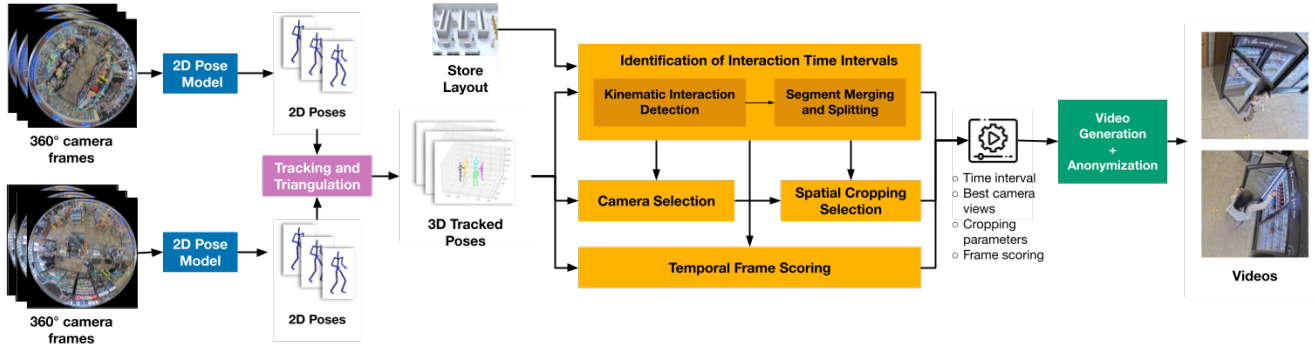


Figure 2. Overview of the RetailAction dataset collection pipeline. The process begins with 360-degree top-view video capture in real-world retail stores, followed by 2D pose estimation and multi-view 3D pose reconstruction. A kinematic interaction detection model identifies potential interaction time intervals, which are refined through segment merging and splitting. The system then selects the best camera views for each interaction, applies spatial cropping, and ranks frames using a temporal scoring algorithm. Finally, videos undergo anonymization and storage, while a two-step annotation process ensures high-quality spatio-temporal interaction labels.

the scope of this paper, whose focus is on the dataset itself.

The automatic video generation system consists of several components. First, the frames from each camera are processed by a model that detects people in the scene and estimates their 2D poses. Since we use 360-degree cameras and traditional 2D pose estimation models do not perform well on such views, we employ a version of the PersonLab model [19], trained on a proprietary dataset composed of 360-degree views and 2D pose annotations collected from multiple stores. The poses generated independently for each camera are then used to perform temporal tracking within each camera view, producing tracklets. Tracklets from different cameras are subsequently clustered based on triangulation error to create unified tracks, which are finally triangulated to obtain 3D position estimates. As a result, we obtain 3D tracks of individuals in the scene along with their 3D poses.

In a second step, we use additional models to identify the time intervals of interest and the best camera views to visualize the events. The first problem is addressed using a kinematic Graph Convolutional Network model based on [23]. This model leverages the 3D poses of each person, along with the positions and dimensions of the store shelves, to detect time intervals with a high probability of containing real interactions of those skeletons with shelf items. Since this model relies solely on 3D pose data and not on images, it does not guarantee that all selected intervals contain actual interactions, but it effectively filters out the vast majority of footage where people are simply walking through the store or standing near a shelf without interacting.

For each time interval, given the positions of the cameras, shelves, and people, we compute a score for each camera based on potential occlusions caused by shelves, and the

visibility of the person and its upper body joints, particularly the hands. Using this score, we select the two best camera views and crop the frames around the person to focus on the interaction while preserving relevant context and minimizing unnecessary background.

Finally, we perform a few post-processing steps to finalize the video generation. First, we apply frame subsampling to reduce the number of frames per video. After this step, each video contains at most 32 frames, regardless of its original duration, resulting in a lower and non-uniform frame rate compared to the original camera footage. Rather than selecting frames uniformly, we employ a model that analyzes the velocity and acceleration of the person’s hands, prioritizing frames with significant movement while down-sampling static segments. Qualitatively, this method has proven highly effective at summarizing long videos or sequences with extended periods of inactivity into a compact set of informative frames, enabling faster video download and contributing to the overall efficiency of the data collection pipeline. Lastly, a post-processing stage applies facial blurring to anonymize all individuals in the videos, and all timestamps are anonymized. Any visible references to store names are either removed or blurred.

4. Data Annotation and Cleaning

To ensure high-quality action labels, we employ a structured two-step annotation process.

Binary classification and data quality labels – First, annotators assign a binary label to each video segment, indicating the presence or absence of an interaction (where interaction implies at least one human-object engagement, while non-interaction includes activities such as browsing or walking). Additionally, they annotate data quality labels

to identify data issues, including:

- *Bad camera view*: failure of the camera selection algorithm, see Section 3).
- *Low resolution*: insufficient resolution to reliably recognize interactions.
- *Few frames*: failure of the dynamic sampling algorithm.
- *Other*: generic pipeline failures such as pose errors.

To enhance dataset quality, at this step we employ a model-in-the-loop strategy: after an initial labeling pass, a model is trained on half of the dataset, and the 10% most disagreeing ground truth–prediction samples are reviewed. This process is repeated three times to refine annotations and improve consistency.

Spatio-temporal fine grained labels – In the second step, annotators precisely mark the temporal boundaries of each individual interaction and spatially localize each interaction in both video views. Annotators are asked to label the point at which the subject’s hand makes contact with the item on the shelf. In videos containing multiple individuals, a red dot is placed on the head of the subject of interest in every frame, ensuring annotators label only the interactions of the designated individual.

Each annotated interaction is categorized into one or more predefined action categories: (1) **Take**, where a subject picks up an item from a shelf, fridge, or counter (e.g., grabbing a sandwich or candy from a shelf); (2) **Put**, where a subject places an item onto a shelf, fridge, or counter after holding it; and (3) **Touch**, where the subject interacts with an item by placing their hand on it without executing either a “Take” or “Put” action. It is important to note that these labels exclusively refer to interactions with retail shelves and are not applied to interactions involving other store entities, such as shopping baskets or checkout interactions.

Data Curation – To ensure the dataset is well-suited as a benchmark for spatio-temporal action localization models, we apply a series of filtering steps following the annotation process. These steps aim to enhance data quality and maintain consistency across multi-view interactions. Specifically, we:

- *Removed single-view segments*: Instances where only one camera captured an interaction are discarded to ensure multi-view consistency.
- *Filtered low-quality samples*: All instances with data quality issues are excluded.
- *Removed outlier segments*: We remove segments with outlier durations, either in their total length or in the duration of individual actions, to exclude cases that might result from occasional errors in the frame scoring model or the temporal identification of segments.
- *Balanced not-interaction segments distribution*: Segments without interactions naturally dominate retail video data as shoppers frequently browse and wander prior to

making a selection. To ensure a better balance between interactive and non-interactive periods, we artificially include only 10% of segments without any action.

5. Dataset Structure and Statistics

Our dataset is composed of a total of 21,000 samples. Each sample consists of two synchronized videos capturing the same scene from two different camera viewpoints. These views correspond to the two highest-ranked cameras, as determined by our camera selection algorithm described in Section 3. The dataset includes videos from 10 different stores, whose distribution is shown in Figure 3. The distribution illustrates the diversity of the dataset. Along with each sample, there is a metadata file with: sampling scores and timestamps from our frame sampling algorithm for every frame (see Section 3), 2D poses and 2D face positions of the principal subject for every frame, and finally spatio-temporal interaction labels.

The labels include, for each action, the temporal interval in frames and a spatial point in pixel coordinates for both video views in the sample, along with the classification of the action type (“take”, “put”, or “touch”). As shown in Figure 3, the number of actions per segment follows a highly skewed distribution, with the majority of segments containing a single action. Regarding the action type, the dataset is also heavily imbalanced, with the vast majority of actions (97.2%) belonging to the “take” class. This is due to the fact that data collection was performed in real stores without scripted actions or actors. While the balance between samples with no action and those with at least one action was artificially adjusted, the distribution of action types among the labeled interactions was not altered from the original data collection. As such, it reflects real customer behavior. This makes the action type less relevant as a distinguishing characteristic within the dataset, but it remains consistent with real-world data distributions.

We emphasize once again that all videos contain at most 32 frames, and that these frames were selected using a model that assigns a score to each original frame (see Section 3). As a result, the videos have a lower and non-uniform frame rate compared to the original footage. Although the number of frames is fixed, the original duration of the segments from which these frames are extracted varies. Figure 3 shows the distribution of the original duration of the segments and of individual actions within the segments, in seconds.

6. Model

To the best of our knowledge, no existing method in the literature directly supports our specific input-output configuration—namely, taking two temporally synchronized videos from different views as input and predicting spatio-temporal

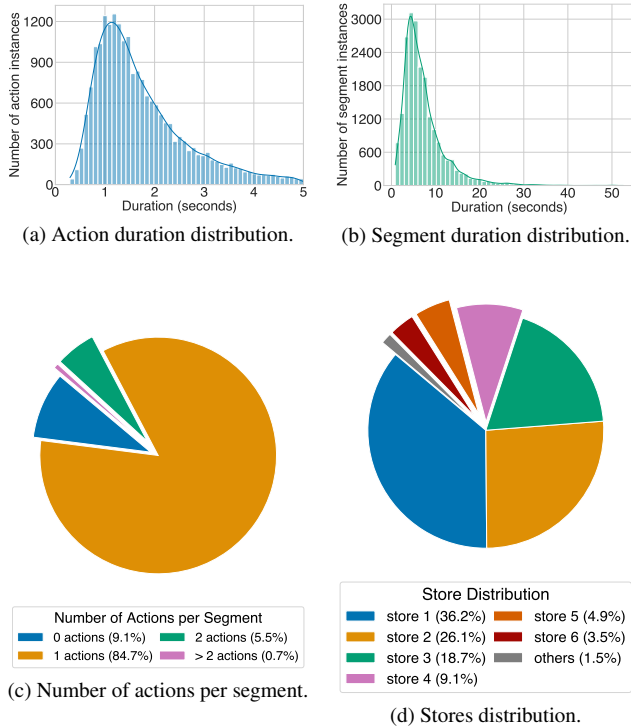


Figure 3. Dataset statistics.

interaction points in both views. In this section, we describe our custom architecture tailored to this task.

Our model takes as input two videos and predicts a list of interactions, each with: the predicted action class, spatial point coordinates for each view, and both the start and end time of the interaction. It consists of different components, as depicted in Figure 4:

Backbone Network – We employ a shared backbone to extract spatio-temporal features from dual-view samples. Each video is independently processed, and the resulting features are then fed to a DETR-based transformer head.

Transformer Head – First, a linear projection reduces feature dimensionality. To capture spatio-temporal relationships, learned 3D positional encodings (two spatial dimensions and one temporal) are then applied. A transformer encoder for each video view refines feature representations using self-attention. Encoded features from both views are concatenated and added to a view-specific positional encoding. Finally, the DETR [2] decoder takes the input queries and global features and generates action localization outputs as a tensor of shape $\#queries \times N$ where $N = C + (S \times V) + 2$ with C denoting the number of action classes, $S = 2$ spatial coordinates, $V = 2$ views, and the last two values representing the action’s start and end times.

Loss Function – The groundtruth-prediction matching

logic builds on the original DETR with the following changes: We predict the spatial coordinates for both views, and use the sum of their distances from the ground truth as one of the factors of the final cost matrix. The other factors are composed of the classification error and the IOU of the temporal predictions. Once assignments are established, the model optimizes classification loss (cross-entropy on predicted action classes), spatial loss (sum of the L2 loss on expected positions for each view), and temporal loss (L1 loss on action start and end times).

7. Metrics

Given that the formulation of the spatio-temporal action localization problem in multi-view retail scenarios is relatively novel compared to conventional spatial and/or temporal localization benchmarks, we have designed an experimental setup specifically tailored to this setting. A key novelty of our dataset lies in the annotation methodology: instead of annotating actions with bounding boxes around the acting person, we provide point-wise annotations indicating the precise location where the interaction occurs.

The standard evaluation methodology in literature benchmarks [15, 21] involves computing the Intersection over Union (IoU) between predicted and ground-truth bounding boxes and subsequently calculating the mean Average Precision (mAP) following COCO or Pascal VOC conventions. In contrast, our task requires predicting a 2D point (x, y) on the image plane for each camera view and associating them with a temporal range $[t_s, t_e]$.

Matching Criteria – Our evaluation protocol follows a COCO-style detection methodology but is adapted to the unique characteristics of our problem, incorporating both spatial and temporal dimensions. A prediction includes a temporal range $[t_s, t_e]$, and a set of spatial points (x_i, y_i) for each camera view i . Assuming a two-view setup, a prediction is represented as:

$$\hat{P} = (c, [\hat{t}_s, \hat{t}_e], (\hat{x}_1, \hat{y}_1), (\hat{x}_2, \hat{y}_2))$$

where c is the maximum score across predicted classes. To determine matches, we compute the IoU for the temporal dimension and the Euclidean distance for the spatial component. Temporal IoU is defined as:

$$IoU_t = \frac{|G_t \cap \hat{P}_t|}{|G_t \cup \hat{P}_t|}$$

where $G_t = [t_s, t_e]$ is the ground truth time range and $\hat{P}_t = [\hat{t}_s, \hat{t}_e]$ is the predicted time range.

For spatial matching, we compute the Euclidean distance between the predicted and ground-truth interaction points in meters. Since our dataset is captured from top-view cameras, direct person area estimation is unreliable. Instead, we normalize distances by computing a pixels-to-meters conversion factor f_i , which is specific to each camera view

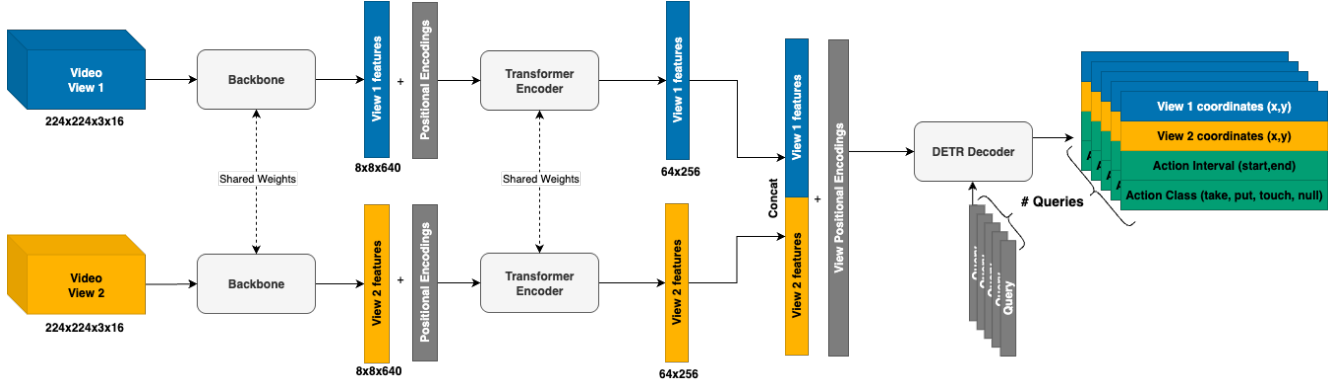


Figure 4. Diagram of the model architecture: Two input video views are processed by a backbone network, generating extracted feature sets for each view. Positional embeddings are added to these features before they are fed into a transformer encoder. The encoder output is concatenated, and a view-specific positional embedding is incorporated before passing it to a DETR-like decoder along with learned action query vectors. For each action query, the model predicts class labels, time intervals, and spatial coordinates for both input views.

i. To estimate this factor, first, we first used a proprietary dataset of over 10,000 3D poses to compute the average length in meters of human bones, provided as part of the metadata. Then, for each video individually, we measured the length of each bone in pixels and divided it by the corresponding average length in meters. The resulting ratios were averaged across all bones of the person in the video to obtain the pixel-to-meter factor for that specific person and video. The spatial distance for view *i* is then given by:

$$d_i = f_i \cdot \sqrt{(\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2}$$

A prediction is considered a match if:

$$\text{IoU}_t > \tau_t \quad \text{and} \quad \min(d_1, d_2) < \tau_s$$

where τ_t and τ_s are predefined temporal and spatial thresholds, respectively. The spatial condition ensures that the prediction is valid if it falls within the acceptable distance in at least one of the available views.

Metric Computation – For each combination of spatial, temporal, and confidence thresholds, we construct a three-dimensional evaluation matrix. Following the standard COCO evaluation protocol, we compute precision-recall curves and enforce monotonicity as in Pascal VOC. From these, we derive two matrices of Average Precision (AP) values: one for spatial thresholds and another for temporal thresholds.

To obtain an overall performance metric, we compute the mean Average Precision (mAP) by averaging over all threshold settings:

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^N AP_i$$

where *N* is the total number of threshold configurations. Additionally, to capture spatial and temporal behavior independently, we compute Spatial Average Precision (mAP_s)

by averaging over the lowest temporal IoU threshold and Temporal Average Precision (mAP_t) over the highest distance threshold. The full curves for different thresholds can also be visualized in Figure 5 to analyze model performance in detail.

8. Experiments

Dataset splits – To ensure a fair evaluation and prevent data leakage, we partition the dataset into three splits, which we release along with the dataset. Since the same individual may appear in multiple videos, we construct the train/validation/test split such that all clips of a given individual appear exclusively in either the training, validation, or test set. Specifically, we allocate 2,000 individuals to the test set, 1,000 to the validation set, and assign the remaining individuals to the training set. This results in 2,501 samples in the test set, 1,277 in the validation set, and 17,222 in the training set. Note that person identifiers are withheld to preserve anonymity. Aside from the individual-based partitioning, all other elements are sampled randomly to maintain a consistent data distribution across splits.

Model Training – For our trainings, we subsample videos to 16 frames and apply preprocessing and augmentation to improve generalization. We apply random cropping and horizontal flipping, with a final cropped resolution of 224 × 224 pixels. The model is trained with a batch size of 10, using gradient accumulation to simulate an effective batch of 1024. We use 5 action queries per sample and train with automatic mixed precision (AMP) on a single A100 GPU. Optimization is done via AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$) at an initial learning rate of 0.0001 with cosine annealing scheduler over 40 epochs. Finally, to improve stability, we apply gradient clipping.

Quantitative Results – To establish baseline performance

for our dataset, we train our head architecture using six different backbones. All models are fine-tuned on the training set and evaluated on the test set, with the exception of ViT-giant [22], for which only the head was trained while the backbone remained frozen. Our evaluation includes two convolutional video backbones—MoViNet-A2 [14] and SlowFast-R101 [8]—as well as four state-of-the-art transformer-based architectures: Multiscale Vision Transformer (MVIT) [7] and three variants of ViT [5] (small, medium, and giant). For all ViT models we use models pretrained using MAEv2 methodology [22] on the Kinetics-700 dataset [4].

Table 2 presents a comparative analysis of all models based on three key metrics mAP, mAP_s and mAP_t . Among the tested models, MVIT-b [7] achieved the highest global and spatial performance, while MoViNet-A2 [14] demonstrated superior temporal localization accuracy. Figure 5 further illustrates the mAP_s and mAP_t curves as functions of spatial and temporal thresholds.

Notably, while model rankings remain relatively stable across temporal thresholds, spatial performance varies significantly across different spatial threshold regimes. For instance, MoViNet-A2 exhibits the lowest accuracy at low spatial thresholds but achieves the highest performance at larger thresholds. This suggests that the model struggles with precise spatial localization. Conversely, models like SlowFast-R101 that perform well at lower spatial thresholds but degrade at higher ones appear to have an appropriate spatial output resolution for the task but may suffer from suboptimal detection behavior.

Model	Finet.	Arch	mAP	mAP_s	mAP_t
Movinet-A2 [14]	✓	Conv	33.5	43.8	60.9
SlowFast-R101 [8]	✓	Conv	40.2	50.4	53.2
MViT-b [7]	✓	Transf.	41.7	55.6	58.2
ViT-small [5, 22]	✓	Transf.	28.3	42.4	46.9
ViT-base [5, 22]	✓	Transf.	31.1	45.7	47.0
ViT-giant [5, 22]	-	Transf.	38.5	50.3	58.0

Table 2. Performance comparison of different backbones. Top two lines are convolutional architectures whereas the others are transformers. Every backbone is pretrained on generic action recognition datasets. ViT-giant backbone has been frozed during finetuning on our dataset.

9. Conclusions

This paper introduced *RetailAction*, a novel dataset and benchmark for multi-view spatio-temporal localization of human-object interactions in real-world retail environments. Unlike existing datasets, RetailAction provides a large-scale collection of videos from operational convenience stores with fine-grained point-based annotations of interactions. With 21,000 samples and 41 hours of video,

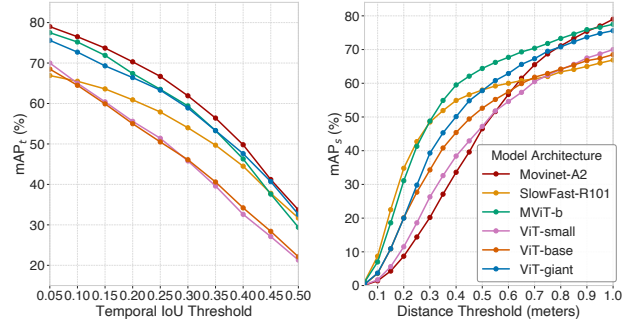


Figure 5. Comparison of mAP_t and mAP_s metrics against temporal IoU and spatial thresholds, respectively. Each curve represents the performance of a different backbone tested.

it presents a more realistic and challenging benchmark for action detection in a retail environment.

Our key contributions include: (i) an automated data collection pipeline leveraging 3D pose tracking, kinematic interaction detection, and dynamic frame rate adjustment; (ii) a precise annotation protocol focusing on interaction locations; (iii) novel evaluation metrics that account for both spatial and temporal accuracy; (iv) baseline performance evaluation using a DETR-based model with state-of-the-art backbones; and (v) public release of the dataset.

Experimental results indicate that while current methods achieve reasonable performance, there is ample opportunity for improvement in precise spatial localization. The best-performing model, MVIT-b, reached a mAP of 41.7%, highlighting the need for approaches that better exploit multi-view data and address the complexities of retail scenes. We also note that the choice of backbone significantly influences performance, suggesting potential for tailored architectures.

Future work could extend the dataset to include more unpublished data, as well as explore models that better integrate multi-view, semi-supervised, and contextual information. We believe that RetailAction will serve as a valuable resource for advancing methods to understand human-object interactions in complex, real-world settings, with potential applications in retail analytics and automation.

References

- [1] Bruno Artacho, Austen Groener, Weijian Li, Yin Wang, Ananth Sadanand, Mohsen Malmir, Sean Ma, Shun Miao, and Quanfu Fan. The groceryvision challenge. <https://grocery-vision.github.io/cvpr2024.html>, 2024. Accessed: 5-Mar-2025.
- [2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

- [3] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A Short Note about Kinetics-600. . Comment: Companion to public release of kinetics-600 test set labels.
- [4] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A Short Note on the Kinetics-700 Human Action Dataset. . Comment: added note about dangers of training on k700 and evaluating on k400/k600. arXiv admin note: text overlap with arXiv:1808.01340.
- [5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representation*, 2020.
- [6] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.
- [7] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835, 2021.
- [8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.
- [9] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The” something something” video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.
- [10] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6047–6056, 2018.
- [11] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos “in the wild”. *Computer Vision and Image Understanding*, 155:1–23, 2017.
- [12] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J. Black. Towards Understanding Action Recognition. In *2013 IEEE International Conference on Computer Vision*, pages 3192–3199.
- [13] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics Human Action Video Dataset.
- [14] Dan Kondratyuk, Liangzhe Yuan, Yandong Li, Li Zhang, Mingxing Tan, Matthew Brown, and Boqing Gong. Movinets: Mobile video networks for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16020–16030, 2021.
- [15] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.
- [16] Ang Li, Meghana Thotakuri, David A. Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The AVA-Kinetics Localized Human Actions Video Dataset. Comment: 8 pages, 8 figures.
- [17] Yi Liu, Limin Wang, Yali Wang, Xiao Ma, and Yu Qiao. Fineaction: A fine-grained video dataset for temporal action localization. *IEEE transactions on image processing*, 31: 6937–6950, 2022.
- [18] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019.
- [19] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *Proceedings of the European conference on computer vision (ECCV)*, 2018.
- [20] Bharat Singh, Tim K. Marks, Michael Jones, Oncel Tuzel, and Ming Shao. A Multi-stream Bi-directional Recurrent Neural Network for Fine-Grained Action Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1961–1970. IEEE.
- [21] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild.
- [22] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14549–14560, 2023.
- [23] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, 2018.
- [24] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8668–8678, 2019.