Submitted to *Information Systems Research*
manuscript

Online Appendix: Predicting Consumer In-Store Purchase through Real-Time Video Analytics: An Advanced Computer Vision and Deep Learning Approach

1. Appendix A: Reference Table for Features and Supporting Theories

Table 1 Theoretical Foundations Supporting Feature Extraction

Dimension	Subcategory	Theory Support	References
Product Interaction	Visual Attention Focus	Visual attentiontTheory	Glaholt and Reingold (2011)
		Eye Tracking in marketing	Martinovici et al. (2023), Ursu et al. (2024), Unger et al. (2024)
	Physical Interaction	Search costs and perceived value	Branco et al. (2012)
		Embodied cognition theory	Barsalou (2008)
		Planned vs unplanned purchase	Hui et al. (2013b)
	Social Interaction	Homophily	Aral and Walker (2011)
		Social facilitation	Zajonc (1965)
		Theory on group size	Mills (1958), Slater (1958)
	Non-Product Engagement	Physical constraints on shopping equipment	Cochoy (2008), Bellini and Aiolfi (2017)
		Phone usage	Atalay et al. (2017)
Spatial-Temporal Trajectory	Proximity to Product Display	Purchase consideration	Hui et al. (2013a)
		Store design and product placement	Sigurdsson et al. (2009), Ferreira et al. (2018)
	Movement Path Speed	Purchase consideration and search cost	Seiler and Pinna (2017)
		Subjective perception of time	Brown (1995)
	Directional Changes Complexity	Exploration and variety-seeking	Hui et al. (2013b)
		Information and search cost	Stigler (1961)
		Physical Surroundings	Bitner (1992)
	Time-Based Engagement	Prolonged decision making: Attention and engagement	Wedel and Kannan (2016), Baumeister et al. (1998)
	Prolonged Decision Making: cognitive load	Baucells and Zhao (2019)	
Pose and Motion	Arm Shoulder Movements	Nonverbal communication theory – openness and consideration vs defensive and resistance	Hall et al. (2019), Burgoon (1994)
	Leg Hip Body Posture	Nonverbal communication theory – Interest and attentiveness	Knapp and Hall (2010)
	Hand Positioning Movement	Nonverbal communication theory –Assertiveness; Readiness to act	Mehrabian (1972)
	Head Orientation and Gaze	Nonverbal communication theory – Agreement and understanding	Ekman and Friesen (1976)
Facial Expressions	Facial Emotions	The impact of emotions on judgments, evaluations, and decisions	Williams et al. (2014), Gardner (1985)
		Facial Action Coding System (FACS)	Ekman and Friesen (1976)
	Eye Gaze	Facial Action Coding System (FACS)	Ekman and Friesen (1976)
		Attention and Trust Building	Kleinke (1986)
	Mouth Status	Facial Action Coding System (FACS)	Ekman and Friesen (1976)
Benchmark Feature	Appearances	Contagion effect	Argo et al. (2008)
		Self-representation	Jensen Schau and Gilly (2003), Clark et al. (1999)
	Demographics	Behavioral difference in gender and age	Mitchell and Walsh (2004), Kotler and Armstrong (2021), Kanwal et al. (2022), Otnes and McGrath (2001), Hansen and Jensen (2009)
	Contextual	Mobile targeting	Baker et al. (2014), Zhang and Katona (2012), Luo et al. (2014), Ghose et al. (2019)

2. Appendix B: Feature Extraction from In-Store Video Data — Framework and Implementation Details

We provide the detailed implementation of our video-based feature extraction process in this section. This appendix offers a comprehensive guide to how rich behavioral features are derived from raw in-store video data using our framework. The technical breakdown—covering feature categories, design rationale, and implementation examples—serves as a practical reference for researchers and practitioners aiming to develop real-time consumer analytics systems leveraging video camera data.

2.1. Reconstructing Consumer Shopping Trajectory Using Person Re-ID

Specifically, we adopted SOLIDER to learn general human representations from massive unlabeled human images, which can benefit downstream human-centric tasks to the maximum extent. Unlike existing self-supervised learning methods, SOLIDER utilizes prior knowledge from human images to build pseudo semantic labels and import more semantic information into the learned representation. Moreover, different downstream tasks always require different ratios of semantic information and appearance information, and a single learned representation cannot fit all requirements. To solve this problem, SOLIDER introduces a conditional network with a semantic controller, which can fit different needs of downstream tasks.

It delivers top-tier results across multiple benchmark datasets: On Market 1501, SOLIDER achieves mAP/Rank-1 = 91.6% / 96.1% (without re-ranking) and improves further to 95.3% / 96.6% with re-ranking. On the more challenging MSMT17 dataset, SOLIDER reaches mAP/Rank-1 = 81.5% / 89.2% with re-ranking. In contrast, competitive Re-ID methods like UFFM (Che et al. (2025)) and CLIP ReID (Li et al. (2022)) baselines typically achieve only 92.0% mAP / 96.1% Rank-1 on Market 1501 and 67–68% mAP / 80–83% Rank 1 on MSMT17. These results validate SOLIDER's superior ability to generalize across complex camera views, making it a strong fit for our retail environment where accurate and efficient tracking across camera views is essential.

We also demonstrated two matched consumer trajectories in Figure 2 and Figure 3.

2.2. Major Feature Set 1 - Consumer Spatial-Temporal Trajectory Features

This feature set captures how shoppers navigate the store, revealing shifts in interest, exploration patterns, and engagement levels that are essential for predicting purchase behavior.

To recover consumers' full Spatial-Temporal Trajectory from video, we conduct location trajectories reconstruction by mapping views from different cameras into a unified plane view with GPS-like coordinates. This is challenging for two main reasons. First, each camera captures only a partial footprint of consumers, requiring us to align all camera views in sequence to reconstruct their true trajectory. Second, since consumer coordinates are detected from fixed camera angles, the resulting movement trajectory may not accurately represent the actual path. Observed movement distances can be distorted—either elongated or shortened—depending on the camera angle.

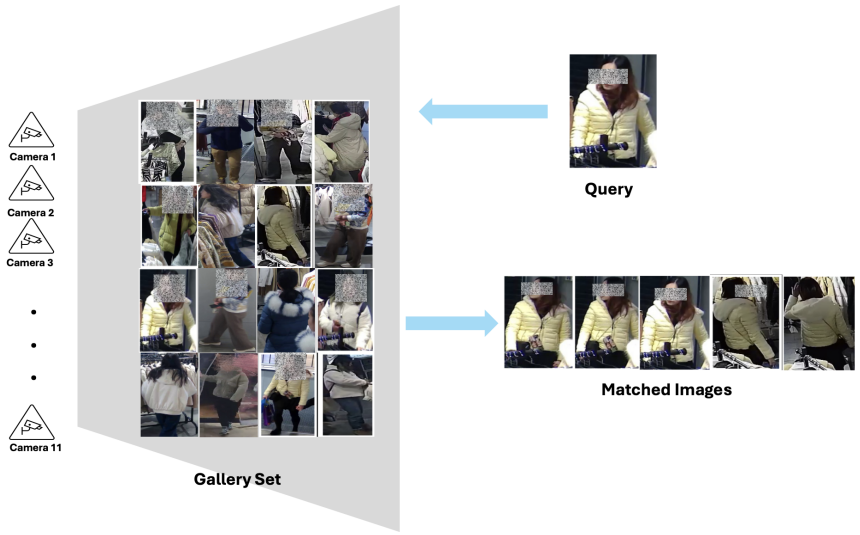


Figure 1 Person Re-ID Pipeline Illustration

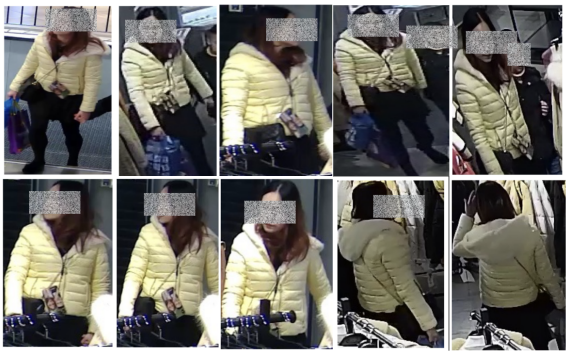


Figure 2 Matched Consumer Trajectories 1

Figure 3 Matched Consumer Trajectories 2

To address these issues, we use homography transformation to map consumer movements across various camera angles, creating a seamless view of the shopping journey. Homography transformation, commonly used in computer vision, maps points from one plane to another under the assumption that both images represent views of the same planar surface.

The process involves determining a homography matrix between two planes by establishing at least four point correspondences—pairs of matching points representing the same physical location on both planes. This homography matrix defines the mapping relationship. Given points $\{(x_i, y_i)\}$ in the camera view and $\{(x_i', y_i')\}$ in the plane view, we estimate the homography matrix $\mathbf{H}$ by solving the following system of equations:

$$\begin{pmatrix} x_i' \\ y_i' \\ 1 \end{pmatrix} \sim \mathbf{H} \begin{pmatrix} x_i \\ y_i \\ 1 \end{pmatrix}$$

This equation shows how each point in the camera view (x_i, y_i) is transformed to a corresponding point in the plane view (x_i', y_i') using $\mathbf{H}$. Solving this system (typically using Singular Value

Decomposition, SVD) provides the homography matrix for mapping any point from the camera to the plane view.

Our goal is to map views from three overview cameras into a single bird's-eye perspective, combining them to reconstruct consumer trajectories in a unified plane view. These overview cameras, capturing the floor from different angles, provide comprehensive coverage of the shopping area. By applying homography transformation, we align each camera's floor plane with the bird's-eye view.

To set up this alignment, we used an overhead map of the mall with precise locations of displays and distances between key points. Ten corresponding ground points were manually labeled in each camera view and on the map. Using these points, we applied OpenCV's homography transformation to estimate $\mathbf{H}$, establishing a one-to-one mapping between each camera's perspective and the plane view. Once the mapping relationships are established, we convert consumer coordinates from each camera into the bird's-eye view coordinates (x', y'). We use consumers' foot coordinates to represent their location in the video view, assuming that the foot position is on the same plane as the ground. To approximate these coordinates, we estimate the midpoint between the left and right foot. The transformation is given by $(x', y', 1)^T = \mathbf{H}(x, y, 1)^T$. By integrating views from all three cameras, we reconstruct the full consumer trajectories in the unified plane view. Example of a converted consumer trajectory in the bird's-eye view is provided in the online Appendix.

Then based on the extracted plane view coordinates, we compute the following set of Spatial-Temporal features:

Proximity to Product Display: This feature measures the distance between a consumer and product displays at each timestamp. To measure this proximity, each product display is modeled as a rectangle with four corners, labeled (x_1, y_1), (x_2, y_2), (x_3, y_3), and (x_4, y_4). We calculate the shortest distance between the consumer's position (p_x, p_y) and the product display by finding the minimum distance to each of the rectangle's four sides. Each side of the display is treated as a line segment between two corner points, such as from (x_1, y_1) to (x_2, y_2).

For each line segment, we calculate the shortest distance from the consumer's position (p_x, p_y) to that segment. We start by representing each line segment as a vector $\vec{v} = (x_2 - x_1, y_2 - y_1)$, and the vector from the consumer's position to the first point of the line segment as $\vec{w} = (p_x - x_1, p_y - y_1)$. The scalar projection parameter t, which determines where the perpendicular from the consumer's position falls relative to the line segment, is given by:

$$t = \frac{\vec{w} \cdot \vec{v}}{\vec{v} \cdot \vec{v}} = \frac{(p_x - x_1)(x_2 - x_1) + (p_y - y_1)(y_2 - y_1)}{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Using the value of t: - If $0 \leq t \leq 1$, the shortest distance from the consumer's position to the line segment is the perpendicular distance; If t falls outside the range $[0, 1]$, the shortest distance

is to the nearest endpoint of the segment. This process is repeated for each of the four sides of the rectangle, and the smallest of these distances is taken as the shortest distance from the consumer to the product display.

Essentially, this shows how close consumers are to displays, helping us gauge interest as they approach certain products. This calculation is performed for every product display and at each timestamp, producing a series of distances that reflect the consumer's proximity to each product display throughout their shopping session.

Movement Path and Speed: This feature set captures consumer movement dynamics by tracking path length and speed over time. Starting with plane view coordinates $(p_{x,t}, p_{y,t})$, which mark the consumer's position on a 2D plane, we calculate both one-step and total path lengths. The one-step path length is Euclidean distance between the current and previous positions.

The total path length, a cumulative measure up to time t, sums all consecutive distances from the start of observation. We also compute one-step speed, representing the consumer's speed between the last two recorded positions, by dividing the one-step path length by the time taken. The average speed up to time t is obtained by dividing the total path length by the elapsed time since the shopping session began.

Directional Changes and Complexity: This feature set measures turning angles, directional changes, and path complexity, providing a comprehensive view of consumer movement patterns within the store. We extract the following features for this group.

- The angle feature measures the change in direction between two consecutive movement vectors. For each time step t, the consumer's position is represented by vectors. The vector $\vec{v}_t$ represents the movement from the consumer's position at time $t-1$ to their position at time t, calculated as $\vec{v}_t = (p_{x,t} - p_{x,t-1}, p_{y,t} - p_{y,t-1})$; Similarly, the vector $\vec{v}_{t-1}$ represents the movement from time $t-2$ to time $t-1$, i.e., $\vec{v}_{t-1} = (p_{x,t-1} - p_{x,t-2}, p_{y,t-1} - p_{y,t-2})$.

The angle θ_t between the two vectors is computed using the dot product:

$$\theta_t = \arccos\left(\frac{\vec{v}_{t-1} \cdot \vec{v}_t}{\|\vec{v}_{t-1}\|\|\vec{v}_t\|}\right)$$

where $\|\vec{v}_{t-1}\|$ and $\|\vec{v}_t\|$ represent the magnitudes of the vectors. This angle helps quantify how sharply the consumer changes direction at each step, providing insights into their navigational behavior.

- The Total Turning Angle represents the cumulative sum of all turning angles made by the consumer over time. This feature captures the total extent to which the consumer has changed direction. A higher Total Turning Angle suggests that the consumer is making frequent or significant directional changes, possibly indicating exploration or indecision.

• The One-Step Direction Change feature identifies whether a significant directional change occurred between two consecutive steps. A direction change is considered significant if the angle θ_t exceeds 45 degrees:

$$\text{one\_step\_direction\_change}_t = \begin{cases} 1, & \text{if } \theta_t > 45^\circ \\ 0, & \text{otherwise} \end{cases}$$

Tracking these sharp changes can highlight moments where the consumer is reconsidering their path or exploring different areas of the store.

• The Total Direction Changes feature counts the cumulative number of significant direction changes up to time t. A direction change is defined as significant when $\theta_t > 45^\circ$. This feature helps in understanding how often the consumer makes meaningful directional changes, which may signal browsing or hesitation.

• The Fractal dimension feature quantifies the complexity of the consumer's path using the box-counting method. The consumer's trajectory is divided into smaller boxes of varying sizes, and the number of boxes needed to cover the path is counted. The Fractal dimension is calculated as:

$$D_f = \frac{\log N(\epsilon)}{\log(1/\epsilon)}$$

where $N(\epsilon)$ is the number of boxes of size ϵ required to cover the path.

Fractal dimension gives us a measure of path complexity: it captures how the consumer's trajectory fills space as the box size changes. A higher Fractal dimension indicates a more intricate, convoluted path, which is typically associated with exploratory or nonlinear movement. In simpler terms, the higher the Fractal dimension, the more complex and meandering the consumer's path, often reflecting more exploratory behavior.

• The Entropy feature captures the unpredictability or randomness in the consumer's movement by analyzing the turning angles at each time step. Entropy is a useful measure for understanding how directed or exploratory a consumer's behavior is within the store. Higher Entropy values indicate more random, less predictable movement, reflecting exploratory or less structured shopping behavior. In contrast, lower Entropy suggests more consistent and deliberate movement, indicative of focused, goal-directed navigation. To calculate the Entropy at each time step t, the turning angles θ_t are first computed for every segment of the trajectory up to time t. From these angles, the probability distribution $P(\theta)$ is determined. The Shannon Entropy H_t up to time t is then given by the formula:

$$H_t = -\sum_{\tau} P(\theta_\tau) \log P(\theta_\tau)$$

where $P(\theta_\tau)$ is the probability of turning angle θ_τ occurring at time step τ.

• The Net Displacement measures the straight-line distance between the consumer's starting point and their current position at time t. Net Displacement reflects how far the consumer has moved from their original point of entry, providing insight into the depth of their engagement with the store environment.

• The Path Tortuosity feature measures how winding the consumer's path is by comparing the total path length to the Net Displacement. It is calculated as:

$$T_t = \frac{L_t}{D_t}$$

where L_t is the total path length up to time t, and D_t is the Net Displacement at time t. A higher tortuosity value indicates a more winding and indirect path, which could suggest the consumer is browsing or uncertain about their decision-making process.

• The average curvature feature measures how sharply a consumer's path bends, indicating whether they follow a straight path (low curvature) or frequently change direction (high curvature), which could suggest browsing or indecision. Curvature K_t at each point is calculated from three consecutive path points, forming a triangle with area A_t:

$$K_t = \frac{4A_t}{a \cdot b \cdot c}$$

where a, b, and c are the triangle's sides. Averaging these values up to time t provides an overall curvature, with higher values reflecting frequent directional changes and lower values indicating a focused path.

The average curvature feature measures how sharply a consumer's path bends as they move through the store. To calculate it, we look at three consecutive points along their path and form a triangle. The sharper the bend, the larger the area of this triangle relative to the distances between the points. If the points form a straight line, the area (and thus the curvature) is zero, indicating no bend. By averaging these curvatures across the entire journey, we get a sense of how winding or direct the consumer's movement has been. A higher average curvature suggests frequent direction changes, hinting at browsing or indecision, while a lower value indicates a more straightforward, focused path.

To calculate curvature K_t at time t, we use three consecutive consumer position points prior to t including $(p_{x,t-2}, p_{y,t-2})$, $(p_{x,t-1}, p_{y,t-1})$, and $(p_{x,t}, p_{y,t})$. The area A_t of the triangle formed by these three points is calculated as:

$$A_t = \frac{1}{2} \left| p_{x,t-2}(p_{y,t-1} - p_{y,t}) + p_{x,t-1}(p_{y,t} - p_{y,t-2}) + p_{x,t}(p_{y,t-2} - p_{y,t-1}) \right|$$

Then we also need to measure the side lengths of the triangle:

$$a = \sqrt{(p_{x,t-1} - p_{x,t-2})^2 + (p_{y,t-1} - p_{y,t-2})^2}$$

$$b = \sqrt{(p_{x,t} - p_{x,t-1})^2 + (p_{y,t} - p_{y,t-1})^2}$$

$$c = \sqrt{(p_{x,t} - p_{x,t-2})^2 + (p_{y,t} - p_{y,t-2})^2}$$

The curvature K_t is calculated as:

$$K_t = \frac{4A_t}{a \cdot b \cdot c}$$

The average curvature up to time t is the cumulative average of curvatures at each step:

$$\text{Average Curvature}_t = \frac{\sum_{\tau=2}^{t} K_\tau}{t-2}$$

The converted GPS-like trajectories from video is shown in Figure 4

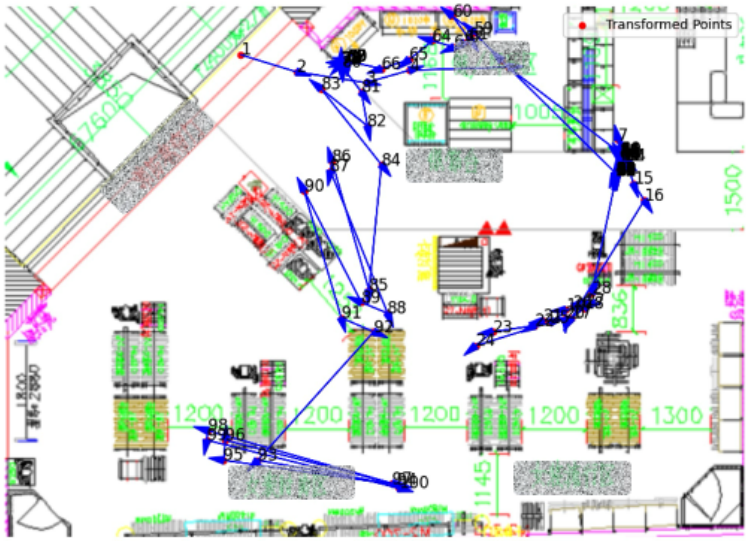


Figure 4 Example of a converted consumers' plane view trajectory

These features collectively offer a detailed view of the consumer's movement within the store. By analyzing directional changes, path complexity, and curvature, retailers can gain valuable insights into consumer behavior, helping optimize store layouts and create targeted marketing interventions to improve the shopping experience.

Time Based Engagement and Dwell Patterns: This feature set tracks how long consumers spend in *dwell areas*, revealing which store sections or products attract the most attention and indicating potential purchase intent.

To identify dwell areas, we use the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm, which detects areas where consumers linger, signaling points of interest. DBSCAN clusters points in a consumer's movement path based on proximity, forming a cluster if at least five points are within a 0.5-meter radius (equivalent to 17.5 units in the coordinate space). If the total time spent in a cluster exceeds 30 seconds, it is classified as a *dwell area*. The following features are then calculated to capture time-based engagement and dwell patterns.

- Time spent up to now measures the total time the consumer has spent in the store from entry up to the current moment. It is calculated by finding the difference between the current timestamp and the store entry time.
- Total dwell time sums the time a consumer spends in all dwell areas. If they remain in a dwell area for more than 30 seconds, that time is added to the cumulative dwell time. This provides insight into how long consumers engage with specific areas, revealing which sections or products attract more attention.
- Belongs to dwell indicates if the consumer is currently in a dwell area. At any moment, it is set to 1 if the consumer is in a cluster and 0 otherwise. This feature helps track real-time engagement, showing when a consumer is actively interacting with products.
- Total dwell count tracks the number of distinct dwell areas the consumer has entered during their shopping session. A higher count indicates multiple moments of high interest, signifying deep engagement with different store sections, which often correlates with a higher likelihood of purchase.
- Revisit dwell measures how often the consumer returns to previously identified dwell areas, which can indicate indecision or re-evaluation of products. Revisiting a dwell area shows continued interest and potential reconsideration of products, suggesting the consumer may be closer to making a purchase decision.

2.3. Major Feature Set 2 - Product Interaction Features

We quantify consumers' interactions with products. This refers to how consumers physically interact with products, such as touching, picking up, comparing, or visually focusing their attention. It also includes social interactions during product engagement, such as showing an item of clothing to another person. These behaviors can indicate how consumers engage with products and reflect varying levels of interest, potentially leading to different purchasing decisions. We categorize different types of product interactions in the following sections.

To capture product interactions from images, we adopt PaliGemma from Google, a versatile and lightweight vision-language model (VLM) inspired by PaLI-3, which is based on open components such as the SigLIP vision model and the Gemma language model. PaliGemma, as a state-of-the-art open-source vision-language model, has been extensively benchmarked and shown to perform

exceptionally well across a range of visual understanding tasks (Beyer et al. (2024)). It takes both image and text as input and generates text as output, supporting multiple languages. It is designed for class-leading fine-tuning performance across a wide range of vision-language tasks such as image and short video captioning, visual question answering, text reading, object detection, and object segmentation. Specifically, we adopted a fine-tuned version for the Visual Question Answering (VQA) task with the VQAv2 dataset[1], google/paligemma-3b-ft-vqav2-224. VQA is a task that involves answering open-ended questions about images, requiring an understanding of vision, language, and common-sense knowledge. We use images as input and ask carefully designed questions for the VQA model to answer.

Physical Interaction: This feature set captures consumers' physical interaction with products, meaning direct contact with the products. We ask the VLM a series of questions to capture different levels of physical interaction, such as: "Is the person touching the product?", "Is the person holding a product?", "Is the person closely examining a product?", "Is the person picking up a product from the rack?", "Is the person trying on a product?", "Is the person comparing two or more products?", "Is the person checking the price tag?", and "Is the person feeling the fabric?" These questions help quantify various types of consumer interaction.

Visual Attention Focus: This type of interaction lacks physical engagement but involves interaction through visual attention. A consumer may be looking at a product or pointing at it, signaling initial interest as the product captures the consumer's attention. We ask VLM model to quantify visual attention through questions like: "Is the person looking at the product?" "Is the person focused on a particular product?" "Is the person looking around without touching the product?" "Is the person browsing through products on the rack?" "Is the person pointing at product items?"

Social Engagement: Another key factor is social engagement during product interactions, such as showing a product to someone or shopping with a companion. We ask the VLM model questions such as: "Is the person showing a product to another person?" "Is the person accompanied by someone?"

Non-product Engagement: We quantify whether the consumer engages in non-product-related activities during product interaction, such as: "Is the person carrying a shopping bag?" "Is the person holding a mobile phone?" "Is the person checking their mobile phone?"

2.4. Major Feature Set 3 - Body Pose and Motion Features

This feature set helps us understand how a consumer moves and behaves in a store environment. These features capture details about the consumer's body position and movement patterns, providing insights into their engagement with products. We quantify consumers' body pose and motion

[1] `https://visualqa.org/index.html`

behaviors by adopting AlphaPose, a state-of-the-art multi-person pose estimator. Pose estimation refers to identifying the position of a person's body parts, like the head, arms, and legs, in a given image or video frame. In simpler terms, it's a way of mapping out where each part of the body is positioned so we can see how someone is standing, reaching, or interacting with an item.

AlphaPose is the first open-source system to achieve over 70 mAP (75 mAP) on the COCO dataset and over 80 mAP (82.1 mAP) on the MPII dataset (Fang et al. 2022, 2017, Li et al. 2019). It detects 136 key full body points, including 26 body keypoints such as the nose, eye, ear, shoulder, elbow, wrist, hip, knee, ankle, head, and neck. Additionally, it identifies 68 face keypoints, 21 left-hand keypoints, and 21 right-hand keypoints. We demonstrate examples of the body points detected for each individual consumer in Figure 5.

Figure 5 Examples of detected body points using AlphaPose

This level of granularity allows us to accurately quantify consumers' movements and actions, such as whether they are leaning forward, raising their hands, bending down, squatting, or tilting their heads. These pose and motion features provide insights into how consumer actions evolve throughout their shopping journey. By extracting 136 key points from each consumer image, we can construct detailed pose and motion features that capture the dynamics of their interactions and behaviors during shopping.

Arm Shoulder Movements: This set of features captures the movements of the arms and shoulders, providing insights into upper body behavior.

- Left and Right Elbow Angle: The elbow angle measures the degree of bending at the elbow joint, providing insights into whether a consumer is reaching for, holding, or manipulating a product. A larger elbow angle (closer to 180 degrees) indicates that the arm is fully extended, often associated with reaching for products, while a smaller angle (closer to 90 degrees) suggests the arm is bent, indicating actions like holding or lifting an item.

$$\theta_{\text{elbow}} = \cos^{-1}\left(\frac{(\mathbf{Shoulder} - \mathbf{Elbow}) \cdot (\mathbf{Wrist} - \mathbf{Elbow})}{\|\mathbf{Shoulder} - \mathbf{Elbow}\|\|\mathbf{Wrist} - \mathbf{Elbow}\|}\right) \tag{1}$$

where *Shoulder*, *Elbow*, and *Wrist* are the respective keypoints.

• Left and Right Shoulder Angle: The shoulder angle measures the arm's position relative to the body. It captures how much the arm is lifted or extended away from the torso. A larger angle indicates that the arm is extended outward, often when the consumer is reaching for or interacting with products, while a smaller shoulder angle suggests the consumer is in a passive state or has completed an interaction.

$$\theta_{\text{shoulder}} = \cos^{-1}\left(\frac{(\mathbf{Neck} - \mathbf{Shoulder}) \cdot (\mathbf{Elbow} - \mathbf{Shoulder})}{\|\mathbf{Neck} - \mathbf{Shoulder}\|\|\mathbf{Elbow} - \mathbf{Shoulder}\|}\right) \tag{2}$$

where *Neck*, *Shoulder*, and *Elbow* are the respective keypoints.

• Upper Arm to Forearm Ratio (Left and Right): This ratio measures the relative extension of the upper arm compared to the forearm. It is calculated by finding the length of the upper arm (the distance between the shoulder and elbow points) and the forearm (the distance between the elbow and wrist points) and then dividing the upper arm length by the forearm length. A higher ratio suggests that the consumer is likely reaching out to engage with a product on a shelf or display, indicating interest and exploration. Conversely, a lower ratio implies that the consumer is holding or closely examining an item, reflecting a deeper level of engagement and a higher likelihood of purchase.

Leg Hip Body Posture: This set of features captures the movements and positioning of the legs, hips, and overall body posture, reflecting lower body dynamics and stability.

• Left and Right Knee Angle: The knee angle measures how much the consumer is bending their legs. A larger knee angle (closer to 180 degrees) indicates standing or walking, while a smaller knee angle (closer to 90 degrees) indicates bending or crouching to inspect lower shelves.

$$\theta_{\text{knee}} = \cos^{-1}\left(\frac{(\mathbf{Hip} - \mathbf{Knee}) \cdot (\mathbf{Ankle} - \mathbf{Knee})}{\|\mathbf{Hip} - \mathbf{Knee}\|\|\mathbf{Ankle} - \mathbf{Knee}\|}\right) \tag{3}$$

where *Hip*, *Knee*, and *Ankle* are the respective keypoints.

• Left and Right Hip Angle: The hip angle measures the degree of bending at the hips. A larger angle indicates an upright stance, while a smaller angle reflects forward bending, typically when examining products at lower levels. A smaller hip angle suggests focused attention on specific items, which may indicate a higher likelihood of purchase. In contrast, a larger hip angle is associated with standing or walking, reflecting more general browsing behavior.

$$\theta_{\text{hip}} = \cos^{-1}\left(\frac{(\mathbf{Neck} - \mathbf{Hip}) \cdot (\mathbf{Knee} - \mathbf{Hip})}{\|\mathbf{Neck} - \mathbf{Hip}\|\|\mathbf{Knee} - \mathbf{Hip}\|}\right) \tag{4}$$

where *Neck*, *Hip*, and *Knee* are the respective keypoints.

• Thigh to Lower Leg Ratio (Left and Right): This ratio captures the relative extension of the thigh compared to the lower leg. It is determined by measuring the distance between the hip and knee (thigh length) and the distance between the knee and ankle (lower leg length), then dividing the thigh length by the lower leg length. A higher ratio suggests that the consumer is standing upright, while a lower ratio indicates that they are bending their knees, which often signifies a crouching or kneeling posture.

Hand Positioning Movement: This set of features captures the positioning and movements of the hands, providing insights into hand-related actions such as reaching, grasping, and manipulating objects.

• Left and Right Hand Position Relative to Shoulder: This feature measures the distance between the hand and the shoulder. A larger distance indicates that the hand is extended outward, reflecting active engagement with products, as the consumer is likely reaching out to grab or inspect an item. Conversely, a smaller distance suggests that the hand is closer to the body, implying that the consumer has picked up the item and is holding it closer for inspection or resting it near the body.

• Left and Right Hand Finger Spread: This feature measures the average distance between the fingers, indicating whether the hand is open or closed. A larger spread reflects an open hand, often associated with exploratory actions, while a smaller spread suggests a closed hand, indicating a firm grip or holding position.

$$\text{finger\_spread} = \frac{1}{N} \sum_{f_1, f_2} \sqrt{(x_{f_1} - x_{f_2})^2 + (y_{f_1} - y_{f_2})^2} \tag{5}$$

where N denotes the total number of pairs of finger keypoints.

Head Orientation Gaze: This set of features captures the orientation and gaze direction of the head, providing insights into where the consumer's attention is focused and how they visually engage with their surroundings.

• Head Orientation (Left and Right): Head orientation measures the consumer's head movement from left to right, indicating where they are looking. This is calculated by measuring the distance between the eye and shoulder points. A larger distance suggests the consumer is looking significantly to the left or right, possibly scanning product displays or navigating the store.

• Head Tilt (Left and Right): Head tilt measures whether the consumer is looking up or down, providing insight into their engagement with products placed outside their direct eye level. It is calculated by comparing the vertical position of the eye relative to the ear. A downward tilt suggests the consumer is looking at products placed lower on shelves, while an upward tilt indicates they are focusing on higher displays.

• Head Facing Direction: This feature captures the direction the consumer's head is facing, indicating which area of the store or display the consumer is focused on. A larger facing direction indicates significant head movement toward a particular area or product.

$$\theta_{\text{facing}} = \cos^{-1}\left(\frac{(\mathbf{MidShoulders} - \mathbf{Nose}) \cdot (\mathbf{MidEyes} - \mathbf{Nose})}{\|\mathbf{MidShoulders} - \mathbf{Nose}\|\|\mathbf{MidEyes} - \mathbf{Nose}\|}\right) \tag{6}$$

where *Mid Shoulders*, *Nose*, and *Mid Eyes* are the respective keypoints.

2.5. Major Feature Set 4 - Facial Emotion, Eye Gaze, Mouth Status Features

We also capture the more granular aspects of consumers' facial dynamics. Capturing facial dynamics, including facial emotion, eye gaze, and mouth status, provides valuable insights into a consumer's emotional and cognitive engagement during the shopping process. These facial cues offer crucial information on the consumer's interaction with products and their overall experience in the store.

Facial Emotions: We use the Face++ API to extract various consumer emotional states, including smile, anger, disgust, fear, happiness, neutral, sadness, surprise. Facial emotions are highly informative for understanding consumer sentiment toward specific products. For example, a smile or expression of happiness may indicate that the consumer is pleased with a product, which can signal a strong likelihood of purchase. Conversely, emotions such as disgust or sadness may indicate dissatisfaction or disengagement, potentially signaling that the product does not meet the consumer's expectations.

Eye Gaze Attention: We use the Face++ API to extract consumers' eye center locations and gaze directions for both the left and right eyes. Each eye is represented by two key aspects: the eye center location and the gaze direction. The eye center location is provided as x and y coordinates, which represent the position of the eye's center on a two-dimensional plane, indicating where the consumer's eyes are located within the frame.

Additionally, we obtain a three-dimensional gaze direction vector consisting of x, y, and z components. The x component of the gaze vector reflects the horizontal direction of the consumer's gaze (left or right), the y component reflects the vertical direction (up or down), and the z component represents the depth or distance, indicating whether the consumer is focusing on something nearby or further away.

These detailed gaze metrics allow us to precisely track where consumers are directing their attention, offering insights into which products or displays are capturing their focus. The longer and more consistently their gaze is directed at a specific area, the greater the likelihood of consumer interest and potential purchase intent.

Mouth Status: We use the Face++ API to detect whether the consumer's mouth is open or closed. This simple yet effective feature can provide valuable insights into consumer behavior. For instance, an open mouth may indicate that the consumer is speaking with a salesperson, asking questions about a product, or expressing interest verbally—activities often associated with a higher likelihood of purchase.

2.6. Major Feature Set 5 - Benchmark Features (Static Features)

In addition to dynamic behavioral data, we also include non-dynamic benchmark features such as consumers' appearance, demographics, and contextual information like the date and time of their store visit. These static features can provide valuable insights into consumer purchase intention by offering additional context about who the consumers are and the conditions under which they shop.

Appearance: We extract consumers' appearance attributes, including beauty level and skin health metrics such as stain, dark circle, and acne. We also determine whether consumers are wearing glasses, categorizing them as dark glasses, normal glasses, or none, using the Face++ API. These appearance features can be insightful in understanding how consumers' self-presentation might influence their shopping behaviors. For instance, consumers who are more appearance-conscious may have different preferences, especially in fashion or beauty products.

Demographic Characteristics: We extract basic demographic information such as gender and age. These demographic features are critical for understanding different consumer segments. For example, younger consumers may have different purchasing patterns compared to older consumers, and gender can often influence product preferences.

Contextual Features: We include contextual features such as the specific day of the week (e.g., Monday, Tuesday) and the time of day (e.g., early morning, late morning, early afternoon, late afternoon, early evening, late evening) when the shopping journey occurs. These contextual factors are important in understanding how consumer behavior varies depending on the timing of their visit. For instance, consumers shopping during late afternoon or evening may be more likely to make immediate purchases, whereas early morning shoppers might be engaging in more exploratory browsing. Additionally, weekend versus weekday shopping behaviors may differ, with consumers potentially spending more time and purchasing more on weekends.

3. Appendix C: Latency Test

To assess whether our framework can operate in real-world real-time contexts, we conducted a real-time feasibility analysis using an actual shopping-day video as a proxy for live input. Specifically, we simulated real-time processing by feeding the Day 13 video (578 consumers) into our framework sequentially as time progressed and measured the total end-to-end latency—from feature extraction to purchase-intent prediction on four NVIDIA A6000 GPUs.

We structured the processing pipeline into three main stages:

1. Step 1 – ReID (Re-Identification) Modules: Person-level tracking to link video frames across space and time.

2. Step 2 – Feature Extraction: Includes parallel processing across four modules: AlphaPose (body pose), PaliGemma (gait dynamics), Spatial-Temporal Trajectory analysis, and Facial/Emotional features (eye gaze, mouth status). Since these extractions are executed in parallel, we take the maximum latency among the four to represent this step.

3. Step 3 – Purchase Intent Prediction: Real-time inference based on extracted behavioral features.

We benchmarked this system under four real-time window scenarios: predicting intent every 30s, 10s, 5s, and 1s. The results are summarized in Table 2 below. The results demonstrate that our pipeline can be confidently deployed in real-time scenarios. When purchase intent is predicted every 30 seconds, the entire end-to-end pipeline—from feature extraction to intent scoring—requires only 0.76 seconds per consumer, allowing nearly 40 users to be processed within each prediction window. Even at higher frequencies, the system remains highly efficient: every 10 seconds requires just 0.42 seconds per user ($\approx$24 users per window), every 5 seconds requires 0.27 seconds per user ($\approx$18 users per window), and even real-time predictions at a 1-second interval can be completed in only 0.12 seconds per user ($\approx$8 users per window).

Table 3 reports the latency breakdown of our end-to-end system under four real-time prediction windows (30s, 10s, 5s, and 1s) for these three steps. Within Step 2, four feature extraction modules—AlphaPose (body pose), PaliGemma (gait dynamics), Spatial-Temporal Trajectory analysis, and Facial/Emotional features (eye gaze, mouth status)—can be processed in parallel. Thus, the effective latency for Step 2 is determined by the maximum runtime among the four modules, rather than their sum.

Although PaliGemma (the vision-language model) is the most computationally intensive submodule within feature extraction (180.5s for 578 consumers at the 30s window), its impact does not undermine overall system performance. With parallelization, the bottleneck shifts to the longest module within Step 2, and the system still delivers end-to-end latency well below each prediction window. Moreover, PaliGemma is served through an optimized VLLM inference engine, which

ensures low-latency serving of generative models. Combined with the modest runtime of other feature extractors (e.g., Pose and Motion, Spatial-Temporal Trajectory, Facial Expressions) and the negligible cost of intent prediction (¡0.4s across scenarios), the overall time cost of our framework remains low.

It is also important to note that our latency measurements are conservative, as they were obtained on four NVIDIA A6000 GPUs without leveraging acceleration toolkits. In practice, significant efficiency gains can be achieved by leveraging tools such as TensorRT[2], TVM[3], or ONNX Runtime[4] for inference acceleration. Additionally, deploying the framework on higher-throughput systems (e.g., NVIDIA H100 or A100 GPUs) can further significantly reduce end-to-end latency, enabling higher throughput and tighter prediction windows. As demonstrated in various NVIDIA performance benchmarks[5], optimized inference pipelines can dramatically increase real-time capacity while lowering system cost per prediction.

Overall, the latency analysis demonstrates that our framework is fully compatible with real-time deployment. By combining efficient consumer trajectory reconstruction, parallelized feature extraction, and low-latency VLM inference, the system enables timely monitoring of dynamic purchase intent, supporting personalized targeting policies that can be executed in real-world retail environments.

Table 2 Latency Test: Processing Time and Capacity by Prediction Interval

Prediction Interval	Time Per Person (sec)	Max Users Processed Per Window
Every 30 seconds	0.76	39.6
Every 10 seconds	0.42	23.9
Every 5 seconds	0.27	18.2
Every 1 second	0.12	8.5

[2] https://developer.nvidia.com/tensorrt

[3] https://tvm.apache.org/

[4] https://onnxruntime.ai/

[5] https://www.nvidia.com/en-us/data-center/resources/mlperf-benchmarks/

Table 3 Real-Time Prediction Latency Breakdown (Time per person, sec)

Step/Module	Submodule/Feature Type	Every 30s	Every 10s	Every 5s	Every 1s
ReID		0.4447	0.2143	0.1277	0.0491
Feature Extraction	Pose & Motion (Alphapose)	0.0597	0.0389	0.0277	0.0143
	Product Interaction (PaliGemma)	0.3123	0.2029	0.1458	0.0683
	Spatial-Temporal Trajectory	0.0065	0.0041	0.0031	0.0015
	Facial Expressions	0.0581	0.0403	0.0318	0.0217
Prediction		0.0007	0.0006	0.0006	0.0006
Total Time per person		0.76	0.42	0.27	0.12

4. Appendix D: Prediction Models and Model Performance

4.1. Prediction Model Architecture

We experiment with four popular deep learning models: LSTMs (Long Short-Term Memory networks), GRUs (Gated Recurrent Units), RNNs (Recurrent Neural Networks) and Transformers. We demonstrate the rationale and details of the model architecture as below.

LSTMs are a type of recurrent neural network that addresses the vanishing gradient problem by introducing memory cells, which help retain information over long sequences. This makes them ideal for modeling consumer actions that unfold over time, especially when certain behaviors earlier in the shopping journey might influence later decisions.

GRUs are a simplified version of LSTMs, using fewer parameters while retaining the ability to handle long-term dependencies in sequential data. GRUs are computationally efficient and are well-suited for scenarios with fewer time steps or when model performance needs to balance between speed and accuracy.

RNNs are the foundational architecture for sequence modeling. However, they are less effective with long-term dependencies compared to LSTMs and GRUs due to issues with gradient propagation, making them more suitable for shorter, less complex sequences.

Finally, we employed a Transformer model (Vaswani et al. (2017)), which represent the state-of-the-art for modeling sequential and time-series data due to their ability to capture long-range dependencies and complex temporal dynamics. Compared with traditional recurrent architectures such as RNNs or LSTMs, the Transformer model offer superior flexibility and performance in modeling consumer trajectories over time (Giuliari et al. (2020), Nie et al. (2022)). By leveraging a self-attention mechanism, the Transformer model dynamically weigh the importance of different time steps, enabling them to effectively capture intricate behavioral patterns and decision-making processes during the shopping journey. We also illustrated the Transformer architecture in Figure 6.

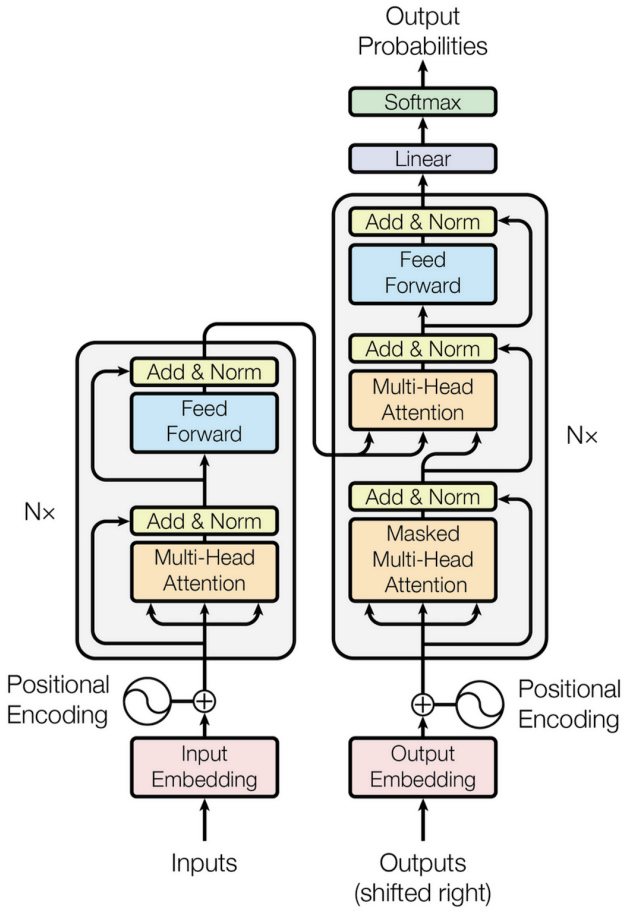


Figure 6 Transformer Model Architecture (Vaswani et al. (2017))

5. Appendix E: Design of Targeting Policy

5.1. Examples of Evolving Purchasing Intention for Purchasers and Nonpurchasers

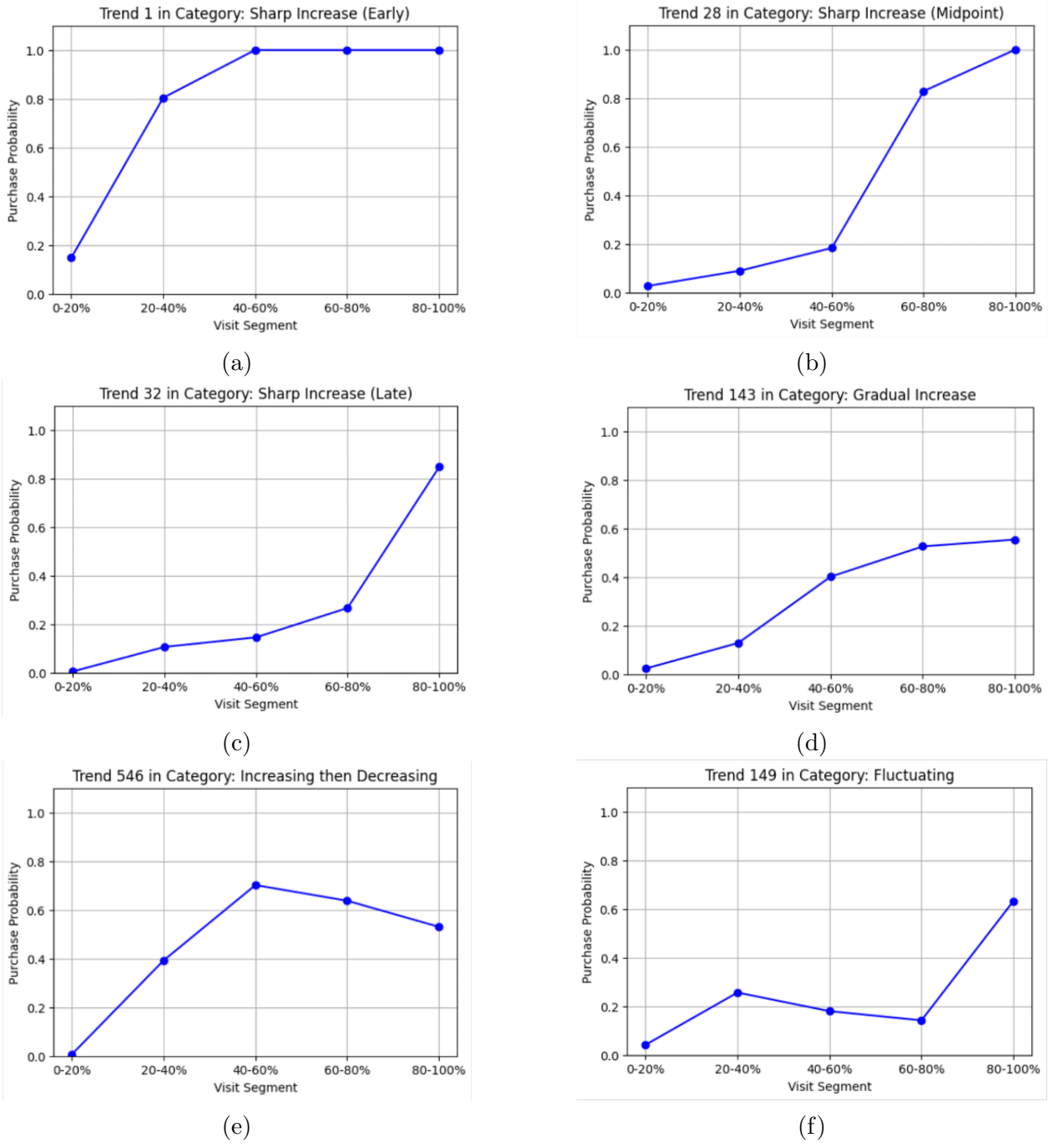
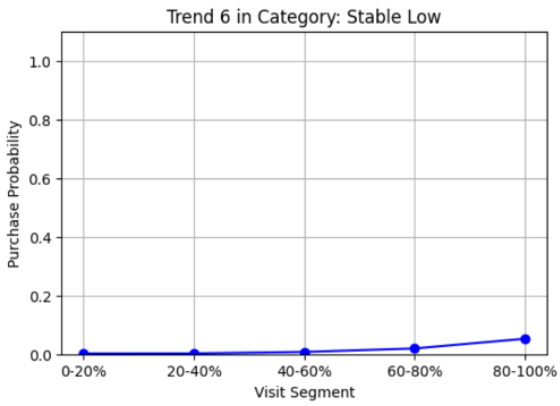
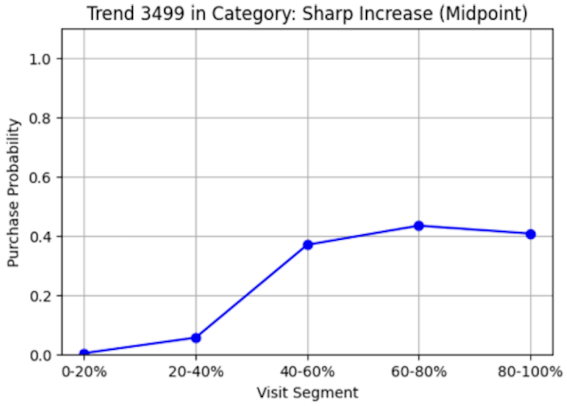


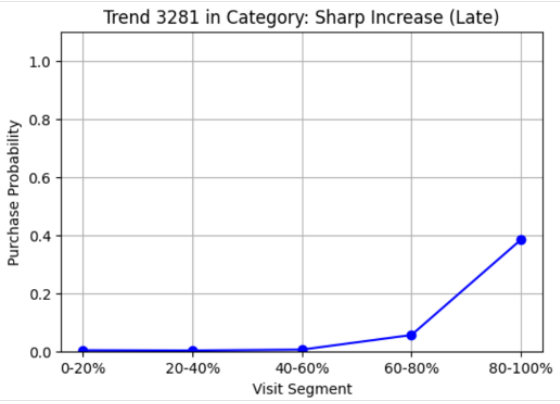
Figure 7 Purchasers Real-time Purchase Probability Examples



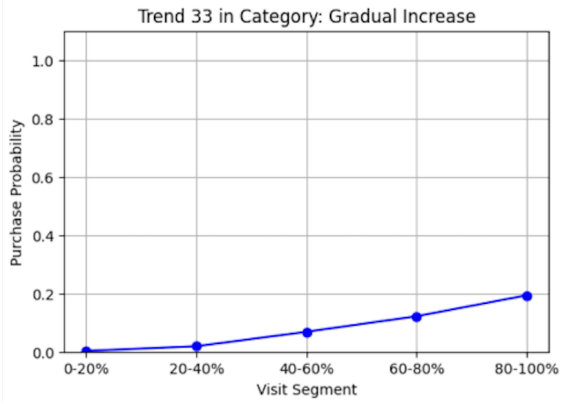
(a)



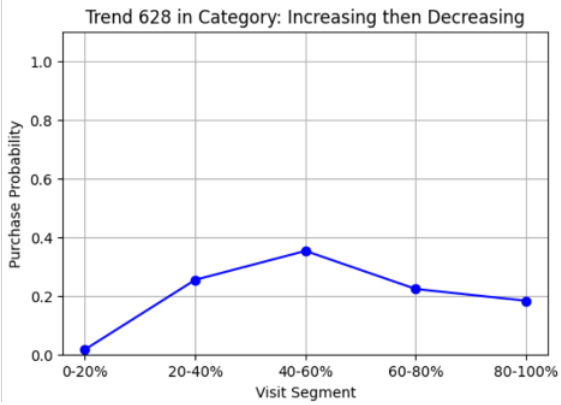
(b)



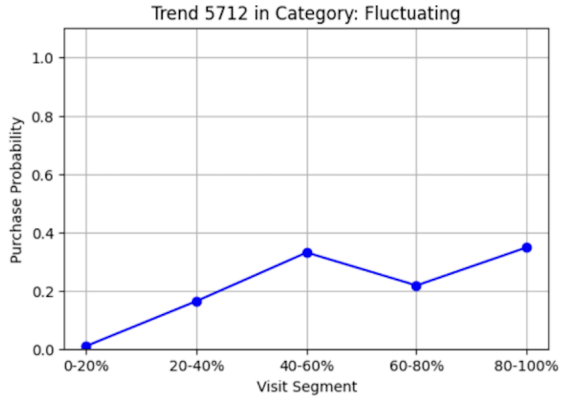
(c)



(d)



(e)



(f)

Figure 8 Non-Purchasers Real-time Purchase Probability Examples

5.2. Simulation Framework for Targeting Policy Lift Analysis

Following Sun et al. (2022), we model the effectiveness of promotional interventions using a probabilistic lift framework. Specifically, for each targeted consumer, a simulated increase in purchase probability is applied at or after the intervention point: $p_{i,t}^{\text{lifted}} = p_{i,t} + \Delta_i, \quad \Delta_i \sim \mathcal{N}(\mu, \sigma^2)$, where Δ_i captures the consumer-specific response to the intervention, drawn from a normal distribution. Consistent with prior work (Ghose and Todri (2016)), which finds that a successful digital ad impression increases consumer intent by an average of 0.22%, we calibrate our simulation using a normal distribution with mean $\mu = 0.0022$ and standard deviation to be $\sigma = 0.22$ (100X mean μ). This setup reflects the heterogeneity in individual responsiveness while preserving the population-level effect size observed in the literature. To reflect real-world marketing constraints, we impose a maximum targeting ratio α, such that at most αN consumers (where N is the total number of shoppers) can be targeted in each scenario. If more than αN consumers qualify, only the top-ranked ones (according to the policy-specific signal) are selected. In our dataset with $N = 12{,}346$ consumers, we set $\alpha = 0.25$, meaning that at most one-quarter of shoppers can be targeted.

The outcome for each consumer is determined by whether their (possibly lifted) purchase probability ever exceeds a conversion threshold (e.g., 0.5). Let $S_{\text{target}} \subseteq \{1, \ldots, N\}$ be the set of consumers chosen for targeting under a given policy, subject to the targeting constraint $|S_{\text{target}}| \leq \alpha N$. For each consumer i, define:

$$\text{Profit}_{\ i} = \begin{cases} R - C, & \text{if } i \in S_{\text{target}} \text{ and } \max_t p_{i,t}^{\text{lifted}} \geq \theta \\ -C, & \text{if } i \in S_{\text{target}} \text{ and } \max_t p_{i,t}^{\text{lifted}} < \theta \\ R, & \text{if } i \notin S_{\text{target}} \text{ and } \max_t p_{i,t} \geq \theta \\ 0, & \text{otherwise} \end{cases}$$

Where

$$\begin{aligned} R &= \text{revenue per purchase} \\ C &= \text{cost per coupon} \\ \theta &= \text{purchase threshold} \\ p_{i,t} &= \text{intrinsic purchase probability} \\ p_{i,t}^{\text{lifted}} &= \text{probability with coupon lift} \end{aligned}$$

Based on our data, the revenue per purchase is set at $R = 67.43$, and the coupon cost is chosen as 1% of the revenue, i,e., $C = 0.6743$. Total profit across all consumers is then $sum_{i=1}^{N}$ Profit $_i$. This simulation procedure allows for a fair comparison of different targeting strategies under consistent assumptions and constraints. While all policies operate within the same structural constraints—such as budget constraint and lift modeling—they differ in the signals used to identify consumers for intervention (e.g., trajectory trends, entropy, hesitation, persuadability zone). The subsequent sections detail the specific design and selection criteria of each policy. We then conduct simulations and use total profit as the key metric of targeting efficiency to evaluate their performance. The pseudocode for the simulation framework is demonstrated in Algorithm 1.

Algorithm 1 General Simulation Framework with Lift Analysis

Input: `consumer_sequences`, `budget_ratio`, `lift_distribution()`, `conversion_threshold`, `revenue_per_purchase`, `cost_per_coupon`, `signal_function()`
for each i in `consumer_sequences` **do**
    $(\texttt{signal}[i],\ t_i^*) \leftarrow \texttt{signal\_function}(\texttt{consumer\_sequences}[i])$
$S_{\text{eligible}} \leftarrow \{i \mid t_i^* \neq \texttt{None}\}$
$N_{\max} \leftarrow \lfloor \texttt{budget\_ratio} \cdot |\texttt{consumer\_sequences}| \rfloor$
$S_{\text{target}} \leftarrow$ top $N_{\max}$ in S_{eligible} sorted by $\texttt{signal}[i]$ (descending)
for each i in `consumer_sequences` **do**
    `lifted_sequence` $\leftarrow$ `consumer_sequences`$[i]$
    if $i \in S_{\text{target}}$ **then**
        `lift` $\leftarrow$ `lift_distribution()`, $\quad t \leftarrow t_i^*$
        for $j = t$ to end **do**
            `lifted_sequence`$[j]$ += `lift`
    `lifted`$[i] \leftarrow$ `lifted_sequence`
for each i in `lifted` **do**
    $m \leftarrow \max(\texttt{lifted}[i])$
    if $i \in S_{\text{target}}$ **then**
        if $m \geq$ `conversion_threshold` **then**
            `profit`$[i] \leftarrow$ `revenue_per_purchase` $-$ `cost_per_coupon`
        else
            `profit`$[i] \leftarrow -$`cost_per_coupon`
    else
        if $m \geq$ `conversion_threshold` **then**
            `profit`$[i] \leftarrow$ `revenue_per_purchase`
        else
            `profit`$[i] \leftarrow 0$
`total_profit` $\leftarrow \sum_i$ `profit`$[i]$
Output: `total_profit`, S_{target}, `profit`

We further evaluate the optimal timing for targeting interventions to maximize total profit. Specifically, we vary the starting point of coupon delivery as a percentage of each consumer's in-store journey—for instance, sending a coupon after 20% or 50% of their shopping duration. The "start fraction" thus represents the share of the consumer journey completed before the intervention. As shown in Figure 9, targeting consumers around the midpoint of their shopping trip (approximately 50%) yields the highest total profit. This finding supports our hypothesis of a timing trade-off in real-time targeting: interventions must balance between gathering sufficient behavioral information from the consumer's early journey and acting promptly enough to capture conversion opportunities before the purchase decision is finalized.

Figure 9 Profit Over Targeting Time (real-time persuadability detection policy)

6. Appendix F: Implementation Guidance for Retailers

In this section, we provide a step-by-step implementation guidance for retailers, along with a system design blueprint for our framework.

6.1. A step-by-step System Design Blueprint

We first provide a modular system design blueprint (Figure 10) that outlines the end-to-end pipeline—from camera input, to feature extraction, to real-time targeting decisions. The modularity allows progressive adoption: smaller stores can start with Modules 1–3 (feature tracking and analytics), while larger retailers can scale into Modules 4–6 (real-time predictive modeling and targeting).

This design outlines a modular, scalable architecture that facilitates end-to-end automation—transforming raw camera input into intelligent marketing interventions—while remaining compliant with privacy regulations and adaptable to varying retail scales.

Module 1: Data Capture Layer (Edge Layer)

Hardware Components: Deploy ceiling-mounted high-resolution RGB cameras across key store zones (e.g., entrance, promotional aisles, checkout). For smaller stores, a minimal configuration with 3–5 strategically placed cameras can still capture high-value zones.

Edge Processing (Optional): For latency-sensitive environments, lightweight edge devices (e.g., NVIDIA Jetson) can preprocess video streams to reduce cloud dependency and preserve privacy.

Module 2: Data Pipeline and Privacy Middleware

Data Streaming: Raw video feeds are streamed securely via an encrypted protocol to a central processing unit (on-premises or cloud).

Privacy Compliance Middleware: Integrate tools for real-time facial blurring, identity obfuscation, and policy-based data retention to ensure GDPR, CCPA/CPRA, or PIPL compliance.
Consent Tracking Subsystem: Log customer consent via app check-ins, loyalty programs, or in-store notices, linked to anonymized session IDs.

Module 3: Perception Layer (Computer Vision Models)
Person Re-identification and Tracking: Use a pre-trained model such as *SOLIDER* to generate consistent anonymized embeddings of individual shoppers across camera zones.
Behavioral Feature Extraction: Apply vision-language models (e.g., *PaliGemma*) to infer fine-grained actions such as product touches, hesitation, and backtracking. Extract features like gaze direction, dwell time, and trajectory curvature in real time.
Temporal Feature Construction: For each consumer, generate a tensor capturing behavior over time, segmented by 1-second intervals.

Module 4: Prediction Engine (Intent Modeling)
Model Architecture: Use a transformer-based model trained on labeled data to estimate time-varying purchase probabilities, denoted as $p_{i,1}, p_{i,2}, ..., p_{i,T}$ for each consumer i using features available up to time t.
Real-Time Inference: At each frame, the model updates the consumer's likelihood to purchase, creating a live behavioral "heatmap" across the store.

Module 5: Policy Layer (Targeting Decision Engine)
Targeting Policy Selection: Choose among five targeting strategies (e.g., high intent, hesitation detection, trajectory volatility) to define S_{target}—the real-time set of customers eligible for intervention.
Coupon Optimization Module: Implement a lift-based rule to avoid targeting customers already highly likely to purchase, maximizing incremental ROI.
Intervention Delivery: Trigger digital coupons via app notifications, shelf screens, or associate prompts depending on store configuration.

Module 6: Monitoring & Simulation Sandbox
Business Simulation Engine: Simulate targeting policies using historical data under varying thresholds and constraints (e.g., budget, coupon cost) to fine-tune rollout parameters.
Performance Dashboard: Provide real-time dashboards tracking targeting accuracy, latency, engagement, and uplift, with KPIs accessible to store managers.

This modular design supports progressive adoption: stores can start with feature tracking and analytics (Modules 1–3), then expand to predictive modeling and real-time targeting (Modules

4–6). The system is horizontally scalable with minimal incremental cost per consumer and enables continuous learning as more video data is collected.

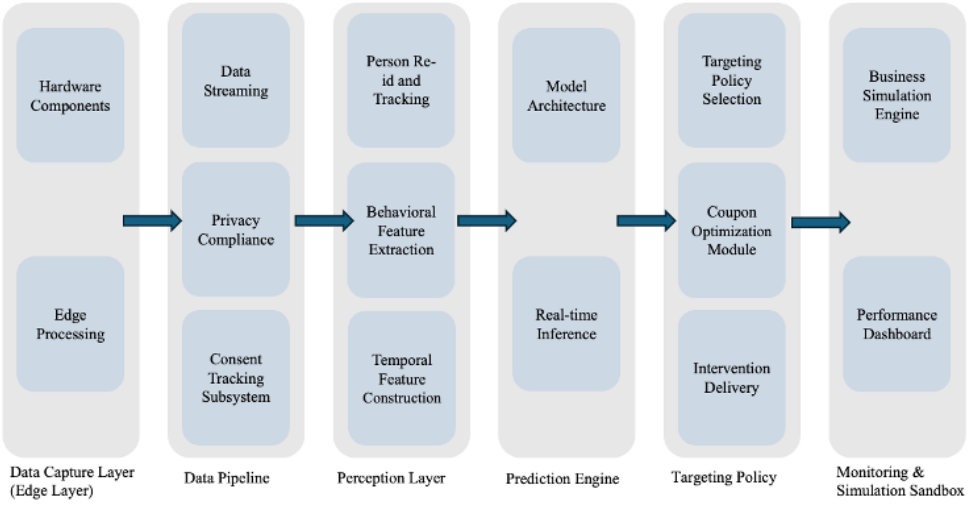


Figure 10 System Design Blueprint

6.2. Implementation Guidance

We then outlined a few practical guidance for retailers regarding actual implementation and potential deployment cost.

- Prediction Frequency and Latency Management: Based on our latency tests, the prediction frequency can be tuned to match store traffic. For example, high-traffic supermarkets may choose shorter windows (e.g., 1–5s) to capture fast-moving shoppers, while smaller specialty stores may use longer windows (10–30s) to balance performance with computational cost. This ensures interventions are delivered at the right moment without overwhelming system resources.

- Targeting Policy Design: Retailers can begin by using our framework to design targeting policies based on real-time purchase-intention trajectories. As shown in our simulations, policies such as the real-time persuadability policy or trajectory-informed policy significantly increase profit compared to baseline approaches. Merchants can select the policy that best aligns with their own operational data and objectives.

- Behavioral Segmentation for Personalization: Drawing on our demographic and trajectory analyses, retailers can translate observed consumer patterns into actionable strategies. For example, male shoppers often follow gradual-intent trajectories, suggesting policies focused on gradual engagement nudges, whereas female shoppers display sharp early increases and hesitation, motivating hesitation-detection interventions at critical decision points. These archetypes enable personalized, context-aware targeting strategies.

- Cost Considerations: The bulk of costs are incurred during initial setup (camera installation, computing infrastructure such as edge devices or servers, and integration with store systems). Once deployed, the marginal cost of processing additional consumers is nearly zero because the system is fully automated, requiring no manual coding or annotation. Smaller retailers can further minimize costs by leveraging existing CCTV infrastructure and low-cost edge devices (e.g., Jetson Nano, Google Coral) to run inference locally.
- Retail environments are inherently complex and dynamic, with factors such as lighting conditions, store layout, camera placement, product assortment, and shopper density evolving over time. These changes can affect the quality and consistency of video-derived signals and may introduce variability in model performance across stores or over longer deployment horizons.

7. Appendix G: Privacy and AI Ethics

We review three representative regulatory frameworks with distinct approaches to video analytics: GDPR in the European Union, PIPL in China, and state-level laws such as CCPA and CPRA in the United States. Despite differences, these frameworks converge on core principles emphasized in international data protection law: transparency and consent (GDPR Art. 7; PIPL Art. 14; CCPA §1798.100), purpose limitation (GDPR Art. 5(1)(b); PIPL Art. 41), data minimization (GDPR Art. 5(1)(c); PIPL Art. 6), data security (GDPR Art. 32; PIPL Ch. IV), and legal/regulatory compliance (GDPR Ch. III VIII; PIPL Ch. VI; CCPA/CPRA §1798.100–1798.199.100). These principles form the foundation for best practices in privacy-compliant deployment of video analytics in retail. Table 4 listed examples of real-world retailers who adopted video analytics practices that are aligned with privacy principles.

Table 4 Examples of Video Analytics Privacy Practices Across Retailers

Principle	Walmart	Kroger	Comcast Smart Solutions
Transparency & Consent	Displays signage in-store ("cameras are operating") and includes privacy clauses in its public Privacy Notice[6].	Kroger prominently uses in-store signage ("cameras in use") and maintains a comprehensive Privacy Notice online[7] that explains data collection, including CCTV.	Promotes "privacy-aware AI"; no personal data used for marketing[8].
Purpose Limitation	Cameras used for loss prevention, safety, and operational insights only[9].	Analytics limited to traffic patterns, fraud/security, and queue management; excludes facial ID[10].	Analytics for safety and insight—not identity tracking[11].
Data Minimization	Footage stored short-term; limited access; no facial recognition[12].	Video processed to extract behavioral metrics then discarded[13].	Partners with Eagle Eye & C2RO; anonymization and no unique ID storage[14].
Data Security	Access limited to authorized staff; encrypted storage; retention schedules[15].	Video encrypted; strict access controls; internal auditing[16].	Cloud storage uses encryption and strict controls[17].
Legal & Regulatory	State law compliance guaranteed; no facial ID; privacy notices in place[18].	CCPA/CPRA aligned; scalable toward GDPR/PIPL[19].	Privacy-aware AI; compliant with local regulations[20].

References

Aral S, Walker D (2011) Identifying social influence in networks using randomized experiments. *IEEE Intelligent Systems* 26(5):62–67, URL http://dx.doi.org/10.1109/MIS.2011.89.

Argo JJ, Dahl DW, Morales AC (2008) Positive consumer contagion: Responses to attractive others in a retail context. *Journal of Marketing Research* 45(6):690–701, ISSN 0022-2437, URL http://dx.doi.org/10.1509/jmkr.45.6.690.

Atalay SA, Bodur HO, Bressoud E (2017) When and how multitasking impacts consumer shopping decisions. *Journal of Retailing* 93(2):1–14.

Baker B, Fang Z, Luo X (2014) Hour-by-hour sales impact of mobile advertising, working paper, Temple University, Philadelphia, PA.

Barsalou LW (2008) Grounded cognition. *Annual Review of Psychology* 59:617–645, URL http://dx.doi.org/10.1146/annurev.psych.59.103006.093639.

Baucells M, Zhao L (2019) It is time to get some rest. *Management Science* 65(4):1717–1734, URL http://dx.doi.org/10.1287/mnsc.2017.3018.

Baumeister RF, Bratslavsky E, Muraven M, Tice DM (1998) Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology* 74(5):1252–1265, URL http://dx.doi.org/10.1037/0022-3514.74.5.1252.

Bellini S, Aiolfi S (2017) The impact of mobile device use on shopper behaviour in store: An empirical research on grocery retailing. *International Business Research* 10(4):58–68.

Beyer L, Steiner A, Pinto AS, Kolesnikov A, Wang X, Salz D, Neumann M, Alabdulmohsin I, Tschannen M, Bugliarello E, Unterthiner T, Keysers D, Koppula S, Liu F, Grycner A, Gritsenko A, Houlsby N, Kumar M, Rong K, Eisenschlos J, Kabra R, Bauer M, Bošnjak M, Chen X, Minderer M, Voigtlaender P, Bica I, Balazevic I, Puigcerver J, Papalampidi P, Henaff O, Xiong X, Soricut R, Harmsen J, Zhai X (2024) Paligemma: A versatile 3b vlm for transfer. URL https://arxiv.org/abs/2407.07726.

Bitner MJ (1992) Servicescapes: The impact of physical surroundings on customers and employees. *Journal of Marketing* 56(2):57–71, URL http://dx.doi.org/10.2307/1252042, accessed 3 Oct. 2024.

Branco F, Sun M, Villas-Boas JM (2012) Optimal search for product information. *Management Science* 58(11):2037–2056.

Brown SW (1995) Time, change, and motion: The effects of stimulus movement on temporal perception. *Perception Psychophysics* 57(1):105–116.

Burgoon JK (1994) Nonverbal signals. Knapp ML, Miller GR, eds., *Handbook of Interpersonal Communication*, 229–285 (Thousand Oaks, CA: Sage), 2nd edition.

Che QH, Nguyen LC, Luu DT, Nguyen VT (2025) Enhancing person re-identification via uncertainty feature fusion method and auto-weighted measure combination. *Knowledge-Based Systems* 307:112737, URL http://dx.doi.org/10.1016/j.knosys.2024.112737.

Clark T, Slama M, Wolfe R (1999) Consumption as self-presentation: A socioanalytic interpretation of mrs. cage. *Journal of Marketing* 63(4):135–138.

Cochoy F (2008) Calculation, qualculation, calqulation: Shopping cart arithmetic, equipped cognition and the clustered consumer. *Marketing Theory* 8(1):15–44.

Ekman P, Friesen WV (1976) Measuring facial movement. *Environmental Psychology & Nonverbal Behavior* 1(1):56–75, URL `http://dx.doi.org/10.1007/BF01115465`.

Fang HS, Li J, Tang H, Xu C, Zhu H, Xiu Y, Li YL, Lu C (2022) Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

Fang HS, Xie S, Tai YW, Lu C (2017) RMPE: Regional multi-person pose estimation. *ICCV*.

Ferreira P, Han Q, Costeira JP (2018) The effect of product placement on shopping behavior at the point of purchase: Evidence from randomized experiment using video tracking in a physical bookstore. *SSRN Electronic Journal* URL `http://dx.doi.org/10.2139/ssrn.3288604`.

Gardner MP (1985) Mood states and consumer research: A critical review. *Journal of Consumer Research* 12(2):281–300.

Ghose A, Li B, Liu S (2019) Mobile targeting using customer trajectory patterns. *Management Science* 65(11):5027–5049, URL `http://dx.doi.org/10.1287/mnsc.2018.3188`.

Ghose A, Todri V (2016) Towards a digital attribution model: Measuring the impact of display advertising on online consumer behavior. *MIS Quarterly* 40(4):889–910.

Giuliari F, Hasan I, Cristani M, Galasso F (2020) Transformer networks for trajectory forecasting. URL `https://arxiv.org/abs/2003.08111`.

Glaholt MG, Reingold EM (2011) Eye movement monitoring as a process tracing methodology in decision making research. *Journal of Neuroscience, Psychology, and Economics* 4(2):125–146.

Hall JA, Horgan TG, Murphy NA (2019) Nonverbal communication. *Annual Review of Psychology* 70(1):271–294, URL `http://dx.doi.org/10.1146/annurev-psych-010418-103145`.

Hansen T, Jensen JM (2009) Shopping orientation and online clothing purchases: The role of gender and purchase situation. *European Journal of Marketing* 43(9):1154–1170, URL `http://dx.doi.org/10.1108/03090560910976410`.

Hui SK, Huang Y, Suher J, Inman JJ (2013a) Deconstructing the 'first moment of truth': Understanding unplanned consideration and purchase conversion using in-store video tracking. *Journal of Marketing Research* 50(4):445–462.

Hui SK, Inman JJ, Huang Y, Suher J (2013b) The effect of in-store travel distance on unplanned spending: Applications to mobile promotion strategies. *Journal of Marketing* 77(2):1–16.

Jensen Schau H, Gilly MC (2003) We are what we post? self-presentation in personal web space. *Journal of Consumer Research* 30(3):385–404.

Kanwal M, Rehman U, Sarwar MI, Khan SU, Jadoon W (2022) Systematic review of gender differences and similarities in online consumers' shopping behavior. *Journal of Consumer Marketing* 39(1):29–43, URL http://dx.doi.org/10.1108/JCM-11-2020-4206.

Kleinke CL (1986) Gaze and eye contact: A research review. *Psychological Bulletin* 100(1):78–100, URL http://dx.doi.org/10.1037/0033-2909.100.1.78.

Knapp M, Hall J (2010) *Nonverbal Communication in Human Interaction* (Wadsworth, Cengage Learning), ISBN 9780495833437, URL https://books.google.com/books?id=om2JPgAACAAJ.

Kotler P, Armstrong G (2021) *Principles of Marketing* (USA: Pearson Education), 18th edition.

Li J, Wang C, Zhu H, Mao Y, Fang HS, Lu C (2019) Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10863–10872.

Li S, Sun L, Li Q (2022) Clip-reid: Exploiting vision-language model for image re-identification without concrete text labels. URL https://arxiv.org/abs/2211.13977.

Luo X, Andrews M, Fang Z, Phang Z (2014) Mobile targeting. *Management Science* 60(7):1738–1756.

Martinovici A, Pieters R, Erdem T (2023) Attention trajectories capture utility accumulation and predict brand choice. *Journal of Marketing Research* 60(4):625–645, URL http://dx.doi.org/10.1177/00222437221141052.

Mehrabian A (1972) *Nonverbal Communication* (Routledge), 1st edition, URL http://dx.doi.org/10.4324/9781351308724.

Mills J (1958) Changes in moral attitudes following temptation. *Journal of Personality* 26:517–531, URL http://dx.doi.org/10.1111/j.1467-6494.1958.tb02349.x.

Mitchell VW, Walsh G (2004) Gender differences in german consumer decision-making styles. *Journal of Consumer Behaviour* 3(4):331–346, URL http://dx.doi.org/10.1002/cb.146.

Nie Y, Chen Z, Wu Y, Liu Y (2022) A time series is worth 64 words: Long-term forecasting with transformers. URL https://arxiv.org/abs/2211.14730.

Otnes C, McGrath MA (2001) Perceptions and realities of male shopping behavior. *Journal of Retailing* 77(1):111–137, URL http://dx.doi.org/10.1016/S0022-4359(00)00047-6.

Seiler S, Pinna F (2017) Estimating search benefits from path-tracking data: Measurement and determinants. *Marketing Science* 36(4):565–589, URL http://dx.doi.org/10.1287/mksc.2017.1026.

Sigurdsson V, Saevarsson H, Foxall G (2009) Brand placement and consumer choice: an in-store experiment. *Journal of Applied Behavior Analysis* 42(3):741–745, URL http://dx.doi.org/10.1901/jaba.2009.42-741.

Slater PE (1958) Contrasting correlates of group size. *Sociometry* 21(2):129–139, URL http://dx.doi.org/10.2307/2785897, accessed 6 Sept. 2024.

Stigler GJ (1961) The economics of information. *Journal of Political Economy* 69(3):213–225, URL `http://www.jstor.org/stable/1829263`, accessed 2 Oct. 2024.

Sun C, Adamopoulos P, Ghose A, Luo X (2022) Predicting stages in omnichannel path to purchase: A deep learning model. *Information Systems Research* 33(2):429–445.

Unger M, Wedel M, Tuzhilin A (2024) Predicting consumer choice from raw eye-movement data using the retina deep learning architecture. *Data Mining and Knowledge Discovery* 38(3):1069–1100.

Ursu RM, Erdem T, Wang Q, Zhang Q (2024) Prior information and consumer search: Evidence from eye tracking. *Management Science* Forthcoming.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Polosukhin I (2017) Attention is all you need. *Advances in Neural Information Processing Systems*, volume 30, 5998–6008.

Wedel M, Kannan PK (2016) Marketing analytics for data-rich environments. *Journal of Marketing* 80(6):97–121, URL `http://dx.doi.org/10.1509/jm.15.0413`.

Williams P, et al. (2014) Emotions and consumer behavior. *Journal of Consumer Research* 40(5):viii–xi, URL `http://dx.doi.org/10.1086/674429`, accessed 8 Sept. 2024.

Zajonc RB (1965) Social facilitation. *Science* 149(3681):269–274, URL `http://dx.doi.org/10.1126/science.149.3681.269`.

Zhang K, Katona Z (2012) Contextual advertising. *Marketing Science* 31(6):980–994.