Shire AI: A Restaurant Intelligence Platform

Automated Table Classification, Fairness-First Routing, and AI Scheduling

Cameron Kuperman Harshith Guduru Alex Tabaku Ben Tang

January 18, 2026

1 Executive Summary

Problem. Restaurant operations traditionally rely on manual monitoring of table states, subjective waiter assignments, and time-consuming schedule creation. These inefficiencies lead to inconsistent customer experiences, unfair tip distribution among staff, and suboptimal resource utilization.

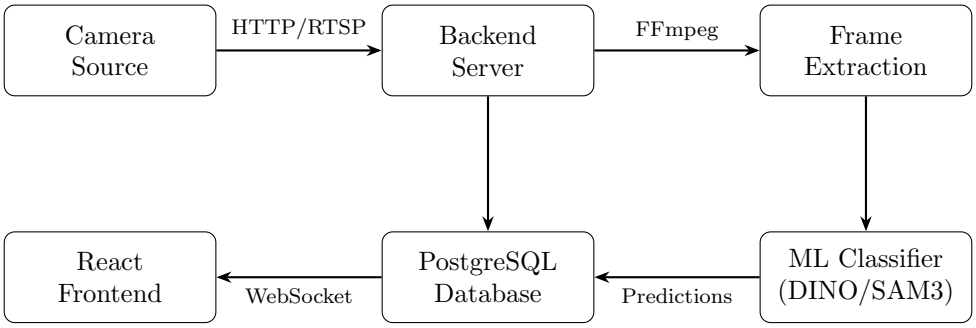
Solution. We present an integrated ML platform that automates three core operational challenges:

1. **Table State Classification**: Real-time CCTV analysis classifies tables as clean, occupied, or dirty using deep learning
2. **Waiter Routing**: Fairness-first algorithm balances workload and tip distribution while maintaining service efficiency
3. **Operational Intelligence**: Data-driven insights for menu pricing, server performance tracking, staff scheduling, and customer review analysis

Key Capabilities. The system processes video streams at >15 FPS, achieving 95%+ classification accuracy with temporal smoothing. The routing algorithm guarantees no waiter receives less than 50% of average tips through an "underserved override" mechanism. The operational intelligence suite provides menu item scoring, Z-score normalized server performance tiers, Gini-optimized scheduling, and LLM-powered review sentiment analysis.

2 System Architecture

2.1 High-Level Data Flow



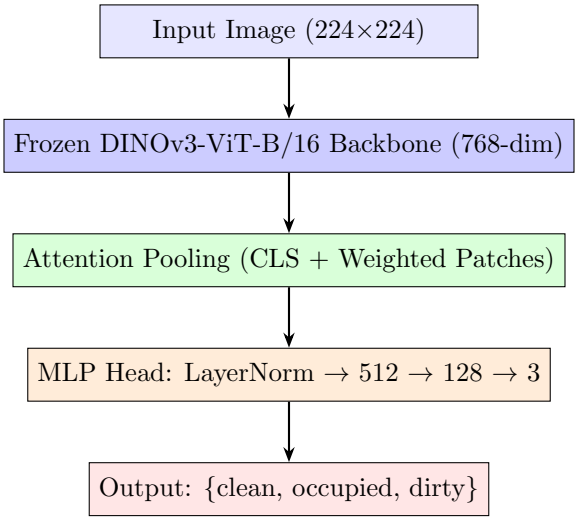
2.2 Technology Stack

Component	Technology	Purpose
Backend API	FastAPI + SQLAlchemy	Async REST endpoints, DB access
Video Processing	FFmpeg + OpenCV	Frame extraction, crop handling
ML Inference	PyTorch + HuggingFace	Table state classification
Database	PostgreSQL	State persistence, audit logging
Frontend	React + Zustand	Real-time floor plan visualization

3 ML Classification Pipeline

3.1 DINOv3 Classifier Architecture

Our primary classifier leverages a frozen DINOv3 Vision Transformer backbone with a custom attention-pooled classification head:



Technical Novelties:

1. **Group-Based Data Splitting**: Consecutive CCTV frames are highly correlated. We group by session+timestamp+table using `GroupShuffleSplit` to ensure train/val/test sets are truly independent, preventing data leakage.
2. **Learned Attention Pooling**: Rather than using only the CLS token, we learn which image patches are discriminative via: Linear(768→128) → Tanh → Linear(128→1) → Softmax. This focuses on plates, dishes, and people rather than background.
3. **Imbalanced Data Handling**: We combine inverse-frequency class weights, Focal Loss $(1-p)^\gamma \cdot \text{CE}$ with $\gamma = 2.0$, WeightedRandomSampler, and Mixup augmentation ($\alpha = 0.2$).

3.2 SAM3 Zero-Shot Alternative

For deployment without training data, we offer a Segment Anything Model 3 classifier using text-prompted segmentation:

```
if detect("person") and mask_area > 10%: return "occupied"
elif detect("plate") and mask_area > 0.5%: return "dirty"
else: return "clean"
```

3.3 Temporal Smoothing

To reduce classification jitter, we implement **N-frame consensus**: states only change when the last N frames (default: 5) all agree on the new classification.

4 Waiter Routing Algorithm

4.1 Fairness-First Scoring

The routing algorithm balances efficiency with fairness using a priority formula:

$$\text{priority} = (\text{efficiency} \times w_e) - \left(\frac{\text{tables}}{\text{max\_tables}} \times w_w\right) - \left(\frac{\text{waiter\_tips}}{\text{total\_tips}} \times w_t\right) - \text{recency} \tag{1}$$

Factor	Weight	Purpose
Efficiency Score	$w_e = 1.0$	Composite: turn time (0.3), tip % (0.4), covers (0.3)
Workload Penalty	$w_w = 3.0$	Prevents overloading any single waiter
Tip Penalty	$w_t = 2.0$	Ensures fair tip distribution across staff
Recency Penalty	1.5	Soft no-double-seat (decays over 5 min)

Underserved Override: If a waiter has <50% of average covers *and* <50% of average tips, the recency penalty is waived. This guarantees no staff member is systematically disadvantaged.

5 Operational Intelligence Suite

5.1 Review Intelligence

Multi-platform scraping feeds ByteDance Seed 1.6 for 5-category sentiment analysis (e.g., food, service), automatically flagging reviews that require manager attention.

5.2 Menu Intelligence

Items are ranked by a composite score:

$$\text{score} = (\text{orders/day})_{\text{norm}} \times 0.5 + \frac{\text{price} - \text{cost}}{\text{price}} \times 0.5 \tag{2}$$

The system suggests +10–15% price hikes for high-demand/low-margin items and flags underperformers (score < 25) for removal.

5.3 Server Intelligence

Waiter performance applies Z-score normalization:

$$\text{performance} = 0.3 \cdot z_{\text{turn\_time}}^{-1} + 0.4 \cdot z_{\text{tip\_pct}} + 0.3 \cdot z_{\text{covers}} \tag{3}$$

Percentile-based tiers ($\geq$p75 Strong, $<$p25 Developing) drive routing priority and LLM-generated improvement insights.

5.4 Scheduling Intelligence

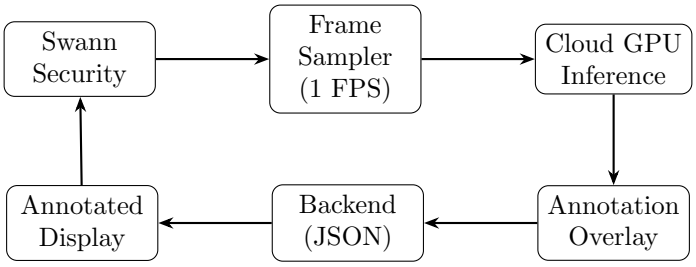
Forecasts use exponential decay (0.85^{weeks}) to feed the assignment algorithm:

$$\text{score} = 0.5 \cdot \text{constraints} + 0.3 \cdot (\text{fairness\_impact} + 50) + 0.2 \cdot \text{preferences} \tag{4}$$

Calculations enforce hard constraints and fairness (Gini < 0.25) while weighting shift preferences. Analytics endpoints support aggregate views and drill-downs.

6 Live Video Integration with Swann Security

For real-time annotation and inference on live camera feeds, we integrated directly with Swann security software:



Pipeline:

1. Swann video feeds are sampled at 1 frame per second
2. Frames are sent to cloud GPU (RunPod) for ML inference
3. Predictions are overlaid back onto the Swann display
4. JSON annotations are persisted to the backend for training data collection

This closed-loop system enables both real-time monitoring and continuous collection of labeled training data from production environments.

Conclusion

This platform addresses core restaurant operational challenges through five integrated systems: ML-powered table classification, fairness-optimized waiter routing, constraint-aware scheduling, review sentiment analysis, and streamlined performance analytics. The combination of computer vision, NLP, and optimization algorithms creates a comprehensive solution that improves both operational efficiency and staff satisfaction.