Leveraging GenAI for Authenticity and De-Duplication in CPG Retail Image Data

Rohit Agarwal
Chief AI Officer
Bizom

BIZOM
business on the move

- **Chief AI Officer** at Bizom
- **18+ years** of Industry Experience
- Started as Software Engineer transformed into Data Scientist about 10 years ago
- Have spoken in **Multiple International Conferences** on Machine Learning and Deep Learning in Europe and USA
- My work on retail analytics is published in all the leading Financial Dailies in India on a regular basis. **In fact Reserve Bank of India (RBI) quotes its consumer index based on our data**
- Currently working on **Generative AI** and its uses for Retail Industry

Agenda

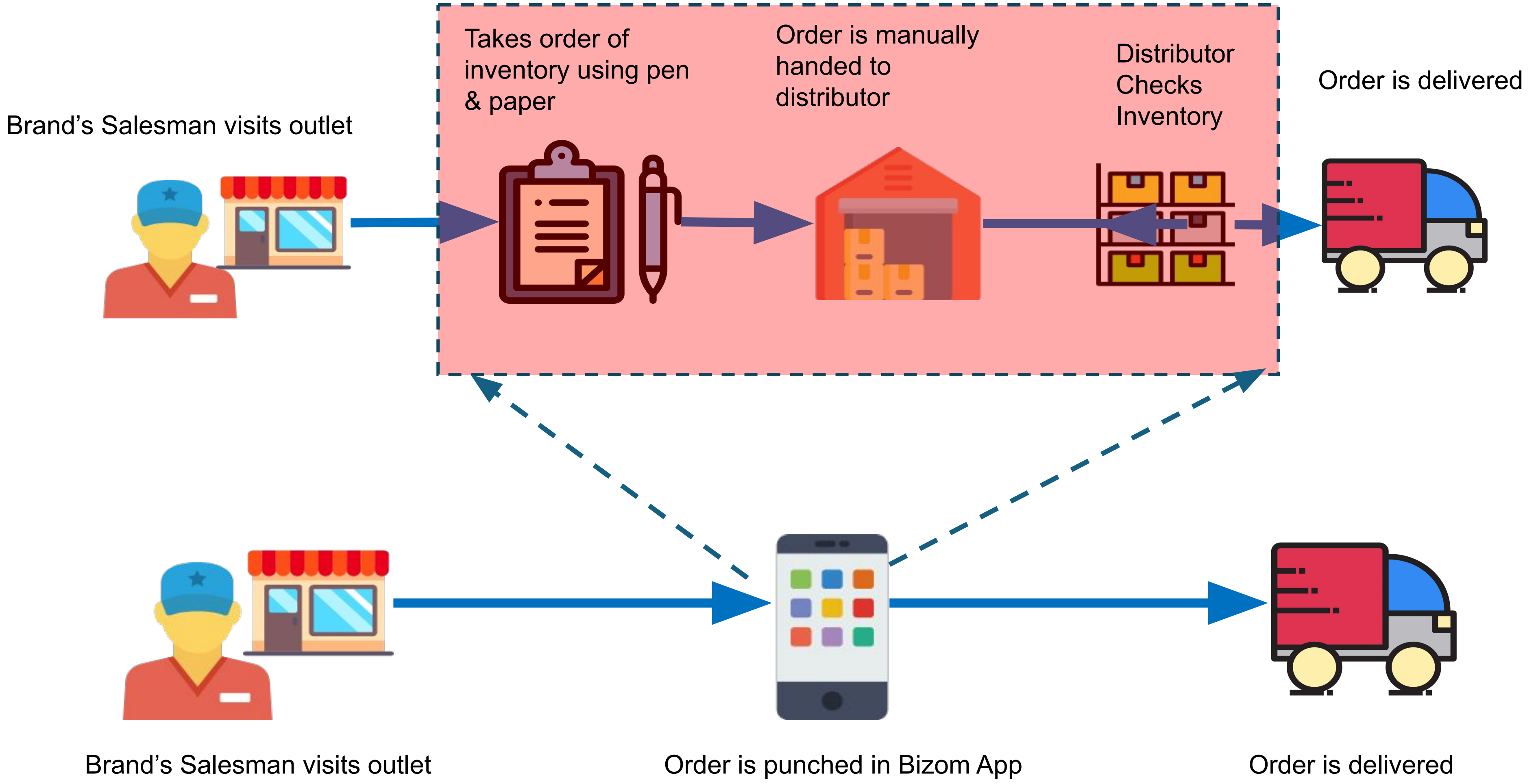
- Who are we?
- Problem Discussion
- Some Solutions
- Case studies from Industry
- Extensions

About us @ Bizom

"Bizom is a Sales Force Automation software that helps in digitizing the steps needed for manufacturers (brands) to place their products into various retail shops/outlets (Kiranas)"

BIZOM
business on the move

Bizom Simplifies Supply Chain



Problem Statement

Retail/Outlet Data

Its Messy!!

Outlet Data

- Name
- Phone No
- Geo Location
- Address

About 10Mn shops look like this !!

The basic information itself is a challenge to get from these shops

BIZOM
business on the move

Why this mess keeps growing?

Practical Issues

- Growth As a KPI for Salesman
  - Fictitious Outlet Creation
- Verification Challenges
  - Audit Difficulties due to volume
  - Lack of Store Signage/ verifiable ID

Impact of the inflated Outlet Universe

BIZOM
business on the move

Inaccurate sales productivity metrics

Distorted sales force deployment

Wasted sales effort & resources

Misaligned targets & incentives

Unreliable reporting

Workaround: Outlet images as part of outlet data

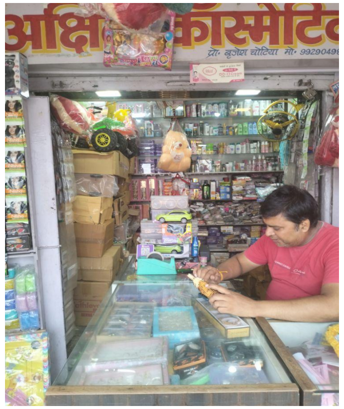
Companies now require a photo along with outlet details

- Intended to reduce duplicate or fictitious entries
- Improves data authenticity at a surface level

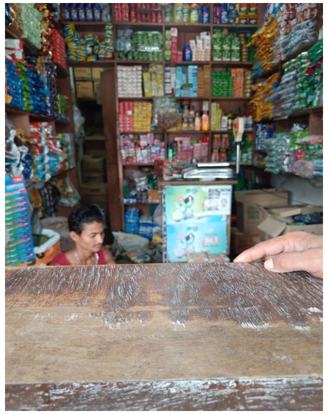
Images can still be reused or manipulated unless further validation measures are in place

FE-2090983

FE-2228133



JA-1026943-2



JA-1032285

FE-1010896

FE-945125

Key Challenges in Outlet Image Verification

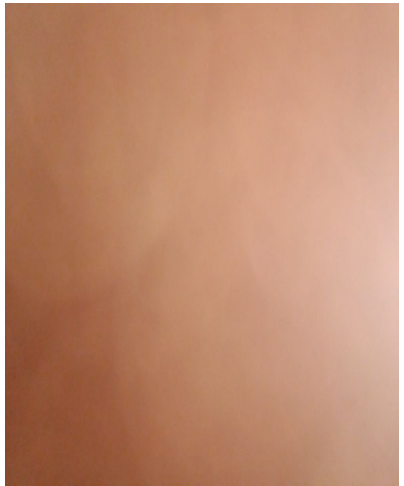
Rather than resolving the issue, the images have only compounded the problem !!!

Problem #1

How to ensure the uploaded image is genuinely of an outlet, and not unrelated content?

Problem #2

How can we identify if the same image is being reused across different outlets?



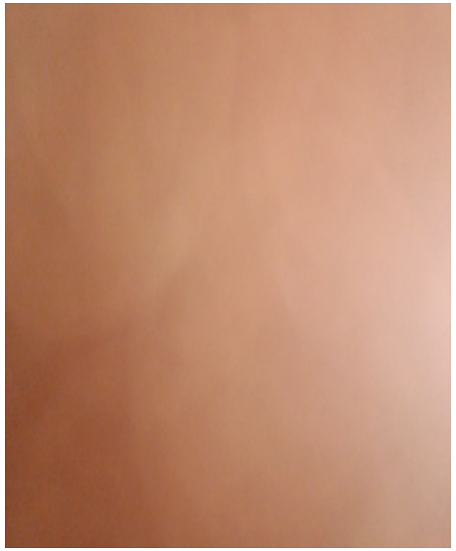
Problem #1: Find out if the image represents an outlet

Attributes of a good outlet image:

- Signage visible
- Products Visible
- Store Open
- Day time photo
- Exterior showing the entire outlet
- ..





BIZOM
business on the move

Why Conventional Object Classification Approach fails?

- 🔄 **High Visual Variability**
  Shops vary widely in appearance—layout, signage, branding—which makes standard classification unreliable.

- 🧠 **Lack of Training Data**
  No consistent dataset exists to train a model that can generalize well across all shop types.

GenAI Characteristics Useful for Retail Domain

Ability to generate data

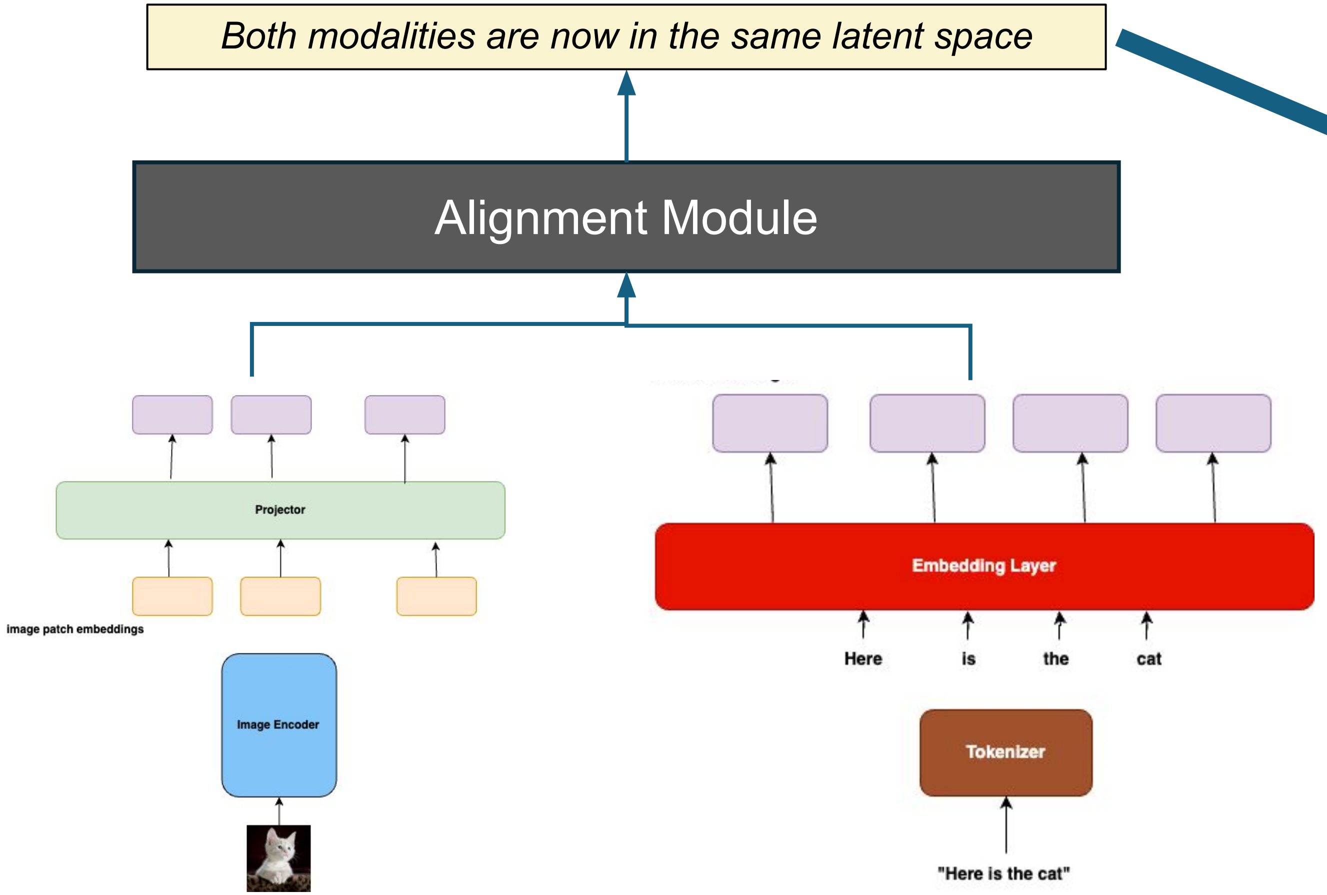
GenAI algorithms like GPT are known to generate text in case of Retail it could be Product Descriptions, Report Summaries

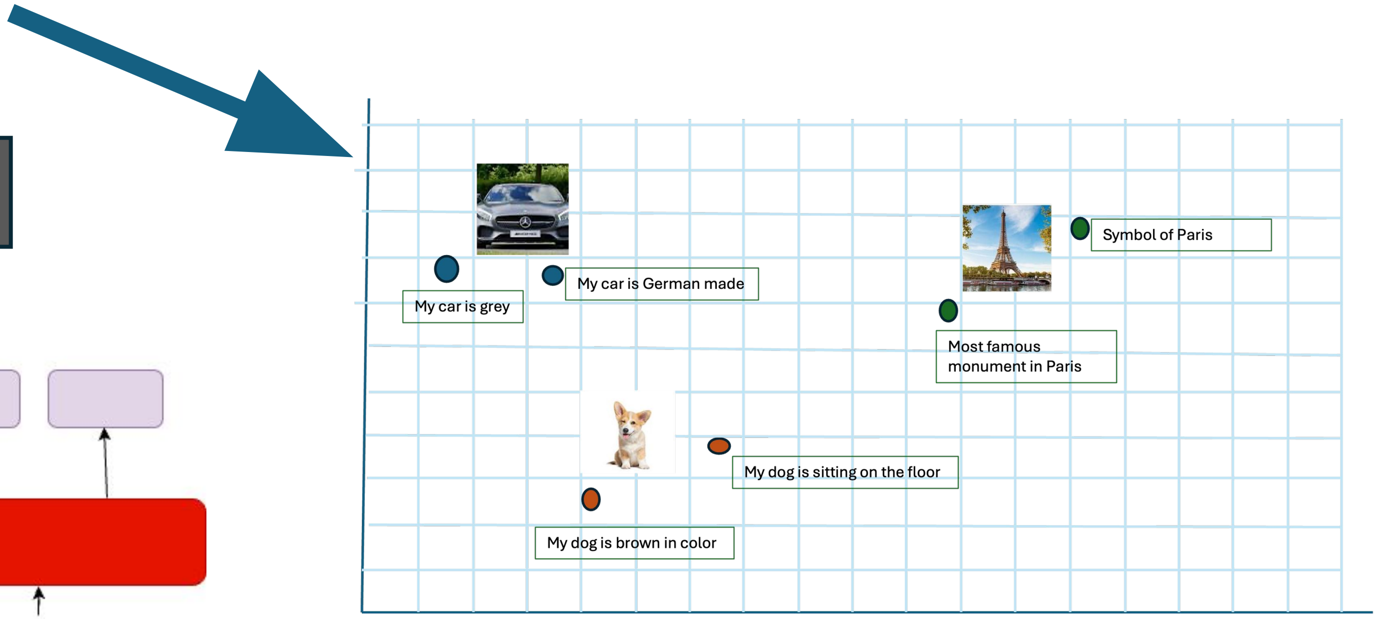
Ability to capture context

GenAI algorithms can analyze text , images to create a mental model of various internal relationships in the data for e.g. how the words in a sentence are related to each other, how different pixels in an image are related etc.

BIZOM

Multimodal LLM





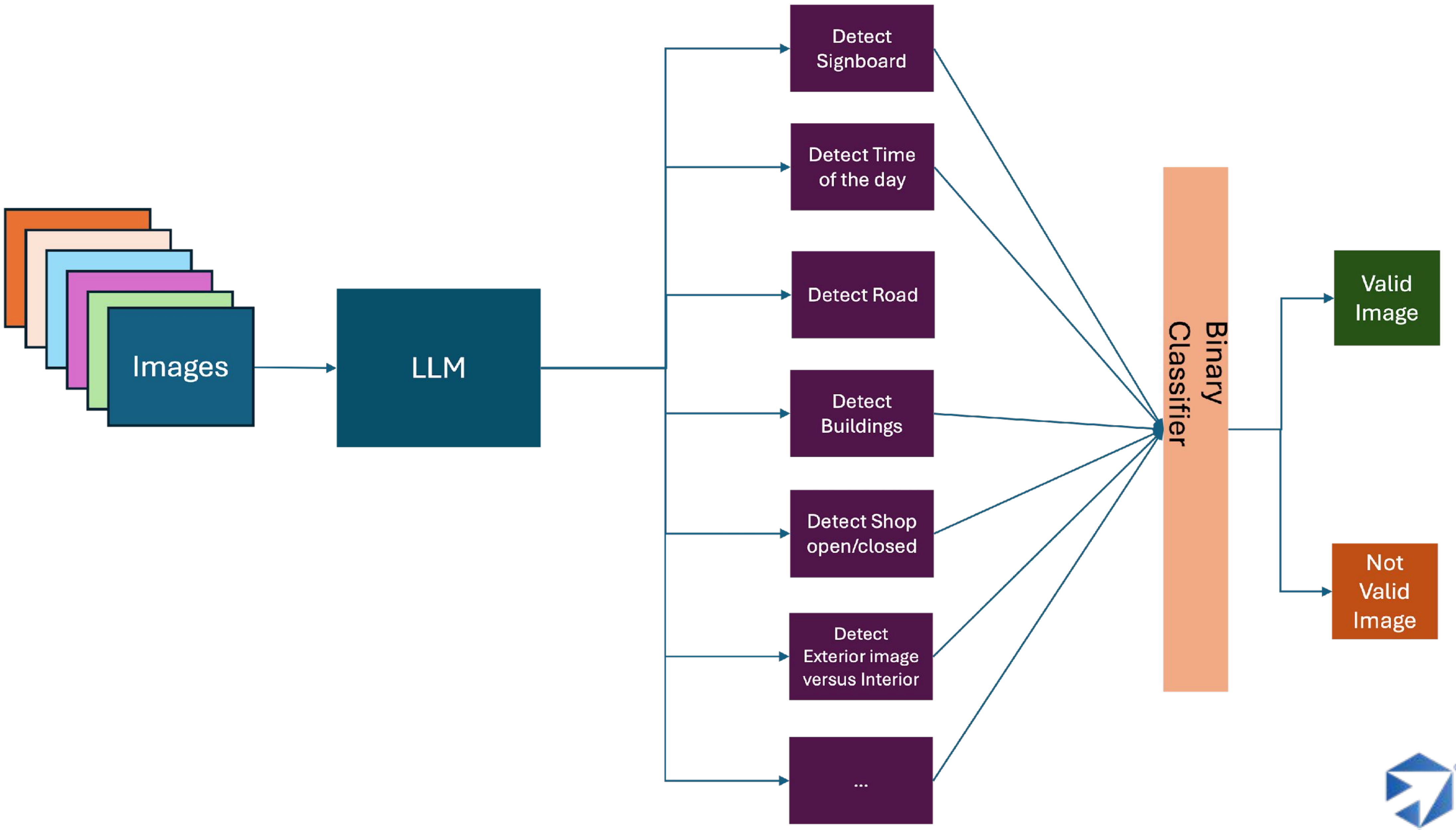
Despite having coming from different modalities, embeddings with similar meaning will be close to each other in vector space

Both the visual and text embeddings can interact now. They can be concatenated together and pass to a LLM to perform some task

https://encord.com/blog/vision-language-models-guide/
https://theaisummer.com/vision-language-models/

BIZOM
business on the move

Approach for finding Image Quality (Pure LLM based)



BIZOM
business on the move

MLLM output on Images

fileName	isHangingProduct	number_people	isShopOpen	number_buildings	isExteriorVisible	isRoadVisible	timeOfDay	isSignVisible	image
outletData/2272629.jpg	no	There are two people near or inside the shop: a man and a woman.	yes	2	exterior	yes	afternoon	yes	
outletData/943147.jpg	yes	There are two people near or inside the shop.	Yes	1	exterior	no	afternoon	no	
outletData/812088.jpg	no	There are no humans visible near or inside the shop.	no	1	exterior	no	afternoon	yes	
outletData/2272626.jpg	no	There are two people near or inside the shop.	no	5	exterior	yes	afternoon	no	
outletData/1777097.jpg	yes	There are two human beings in the shop: a man and a woman.	Yes	1	interior	no	afternoon	no	
outletData/875488.jpg	no	There are no human beings visible in the image.	no	0	exterior	no	night	no	
outletData/2272622.jpg	yes	There is only one person, a man, near or inside the shop.	yes	1	interior	no	afternoon	no	

BIZOM®
business on the move

Issues with pure LLM approach

Inconsistent Accuracy

Highly Dependent on Image Quality

Lack of Local Context

Extremely Slow at Scale

Our Approach: Two stage process

Stage1:
Using MLLMs Generate Pseudo-labels which act as Training Data for image classifier

Stage2:
Train a Vision Model like ResNet or ViT to act as an image classifier

Training Setup

Data Set	#of Images
Training Set	8000
Validation Set	1000
Test Set	1000

Dataset

Outlet Type Variations: Captures Diverse Retail Environments Like Different Shop Types: Medical, Bakeries, Convenience

Regional Variations: Images of Retail Shops from different regions of country

Inherent issues while capturing Images: no Signage, Dark Interiors, People Standing in Front, Shutters Half down

Dataset Images

BIZOM
business on the move

Stage1: Pseudo-Label Generation

Input

Domain Specific Prompt Engineering

Create prompts to understand if the image is of a Kirana store.

Prompt Examples

“Is the retail outlet fully visible with its boundaries?”
“Is the store board clearly visible and not occluded?”

Ensembling

Multimodal Evaluation

The same input (image + prompt) is fed into **three pre-trained MLLMs**

Model Ensembling

Each model independently **assesses alignment** between the image and prompt and makes **Binary Decision Output**

- ✔️ **Positive** – Image aligns with the prompt
- ❌ **Negative** – Image does **not** align with the prompt

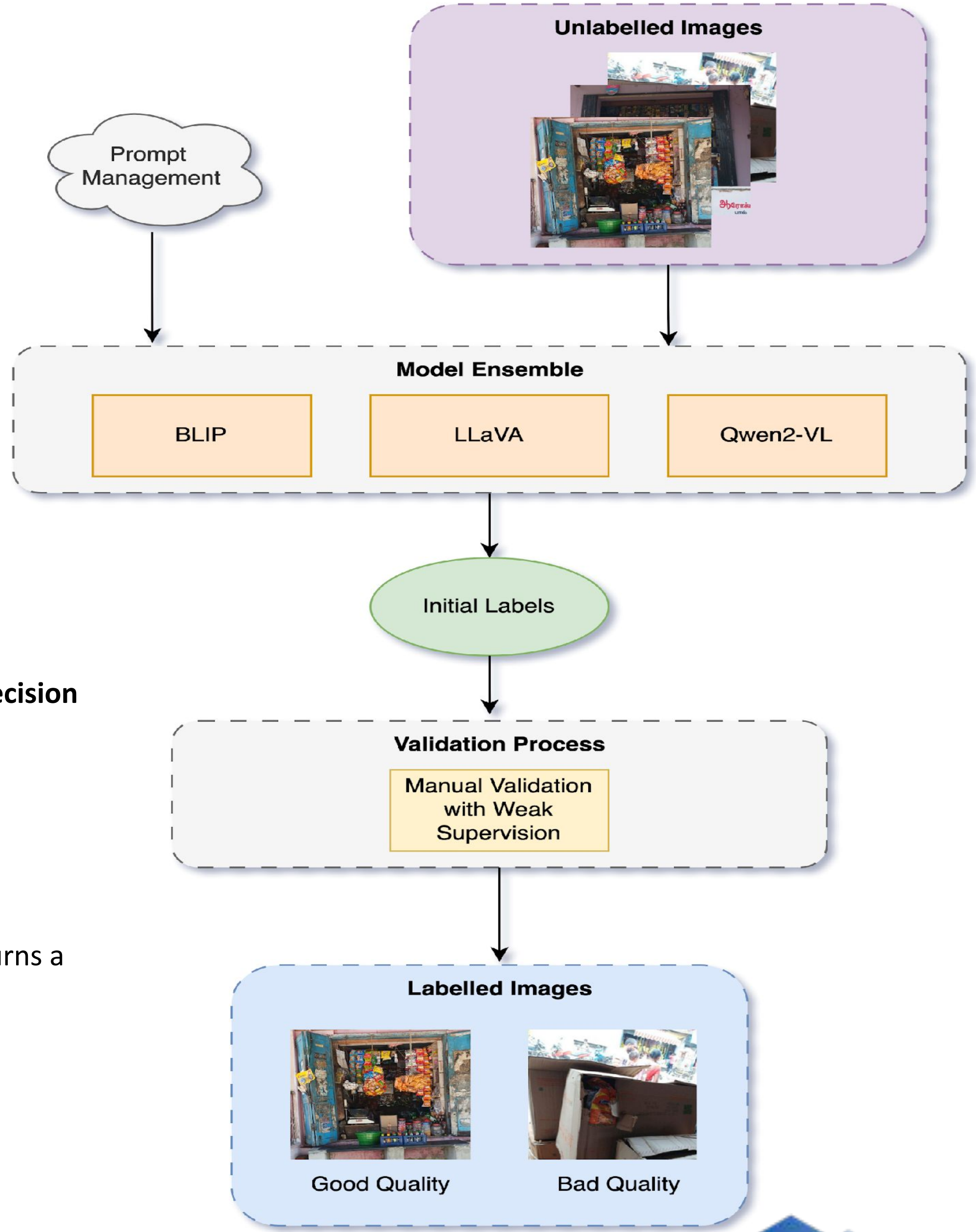
Unanimous Voting

An image is labelled positive only if all three models agree it is of good quality. If even one model returns a negative decision, the image is labelled negative .

Validation

Manual Validation

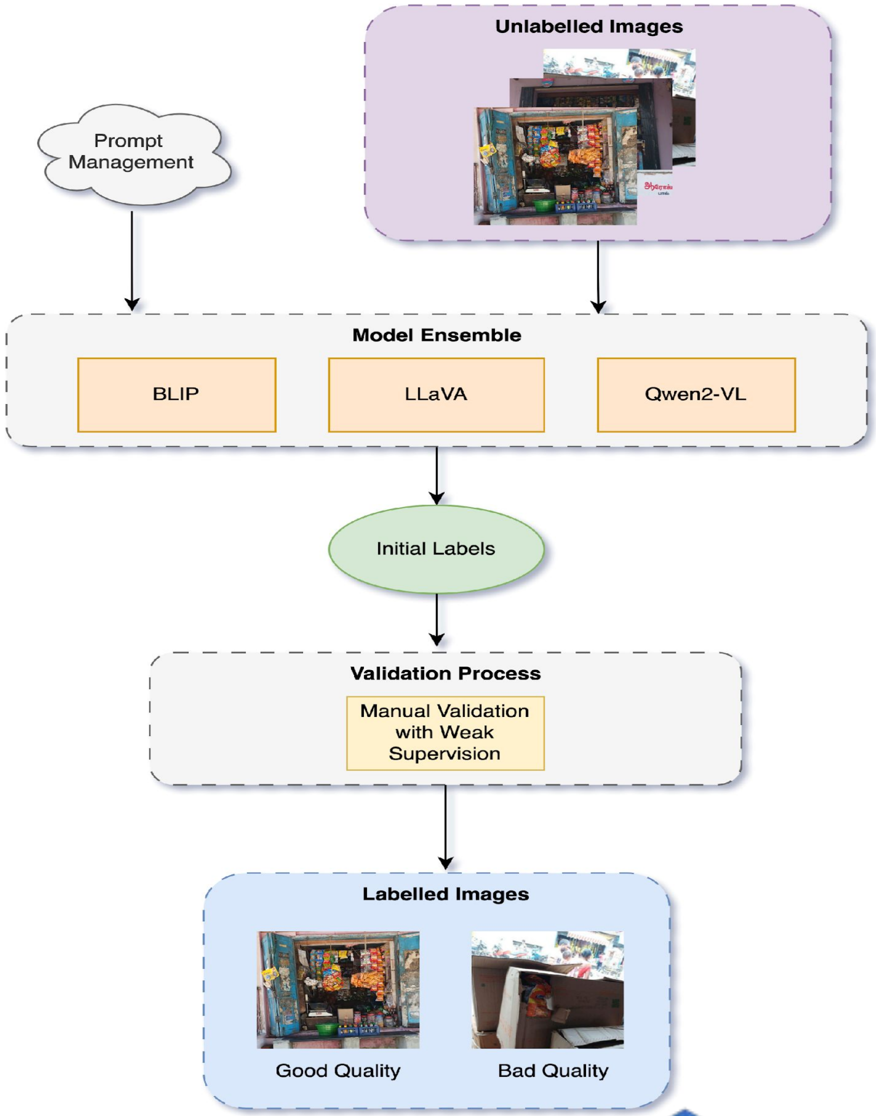
A quick manual validation is done to ensure that we pass only the correctly labelled images to the classifier training stage



BIZOM
business on the move

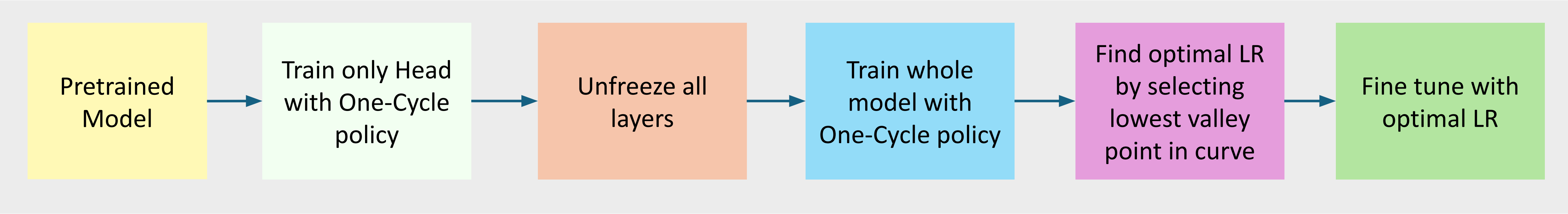
Stage1: Pseudo-Label Generation LLMs contd…

Model	Accuracy	Precision	Recall	F1-score
Blip	0.4869	0.3569	0.9045	0.5118
LLaVA	0.3420	0.3085	**0.9771**	0.4690
Qwen2-VL	0.7635	0.5587	0.9744	0.7102
Our Approach	**0.8066**	**0.6236**	0.8816	**0.7305**



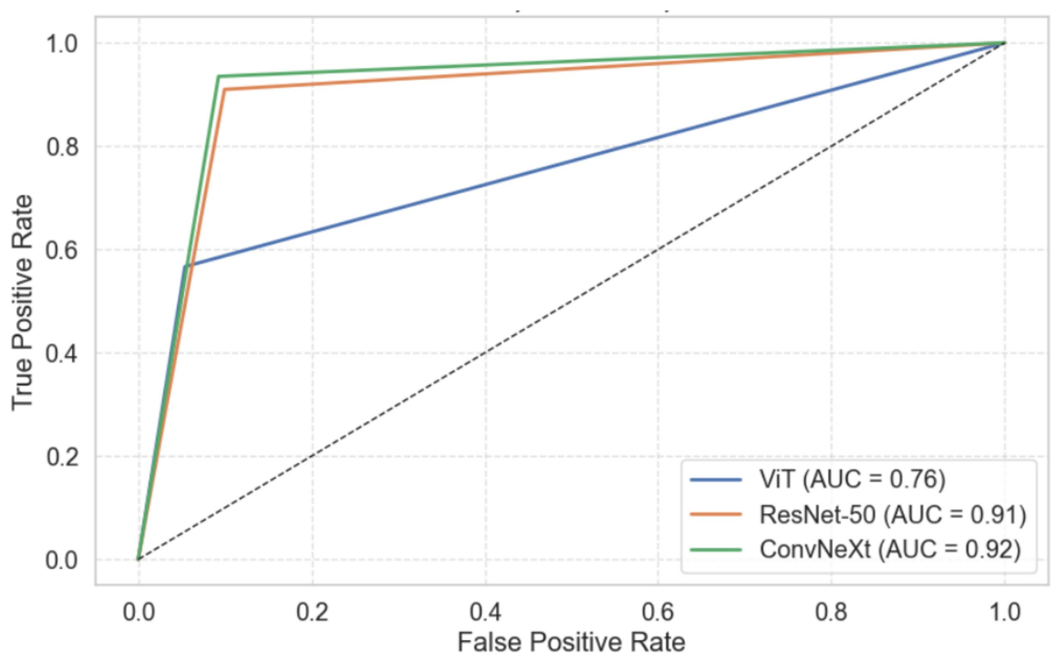
BIZOM
business on the move

Stage2: Training of Vision Classifier



Model	Accuracy	Precision	Recall	F1-score
ConvNext	**0.9150**	**0.9238**	**0.9150**	**0.9170**
ResNet-50	0.9030	0.9117	0.9030	0.9052
ViT	0.8410	0.8370	0.8410	0.8316

- **ConvNeXt** yields the best classification performance but comes with higher computational requirements.
- **ResNet-50** offers a strong balance between performance and efficiency, aligning with real-world deployment constraints.



BIZOM
business on the move

Results

Not Acceptable Images

Acceptable Images

Advantages of the approach

💰 **Cost and computation efficient** solution

✅ Achieves **>90% accuracy** on 1MN outlet images for an enterprise customer

⚙️ **GPU used only for Pseudo Label creation** on a small data sample

🖥️ **Classifier runs entirely on CPU** post-labelling

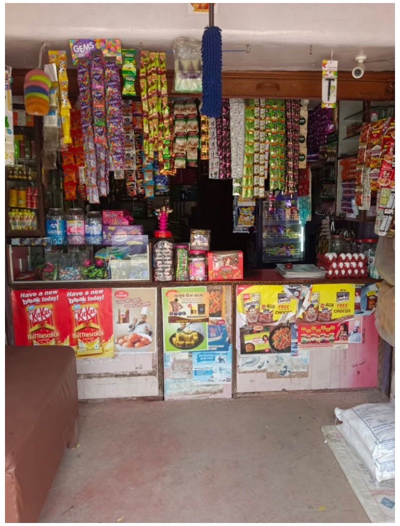
🚀 **10x faster** than GPU-based alternatives

📉 **Significantly reduced manual effort** needed for training data generation

🤖 **Automated pipeline** ensures speed and reliability

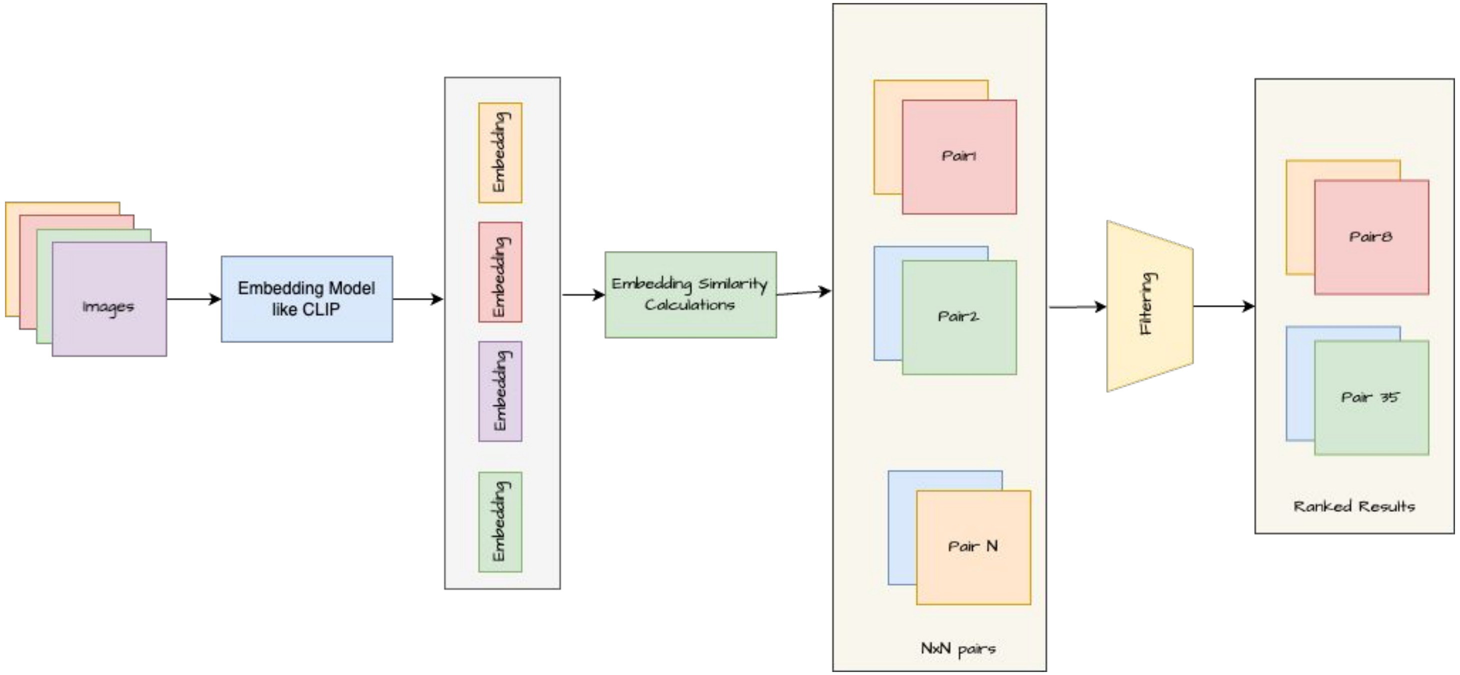
BIZOM
business on the move

Problem #2: Find out if same image has been used for another outlet



Outlet# 7213

Outlet# 6182



- We employ LLM model to create the contextual vectors from the image data
- The Algorithm is quite fast and finding out duplicates among 10K images took less than 1 minute (NxN comparisons)

BIZOM
business on the move

What is an Embedding?

Embedding is the numerical representation of the words (tokens) such that they can be used in the machine learning models. They are also called vectors as they are 1-D arrays of real numbers

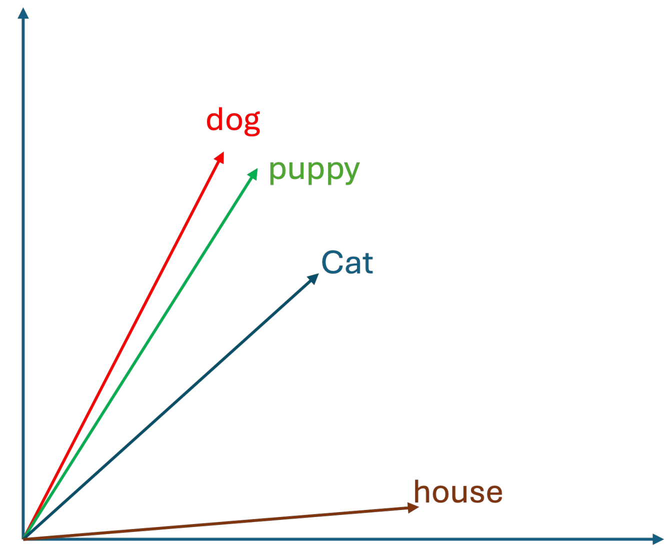
🔢 Embeddings are vectors that allow mathematical operations like computing distances to measure similarity.

🚫 By default, embeddings are independent and lack relational meaning.

🧠 Neural networks enrich embeddings with **contextual understanding**, enabling intelligent language-based workflows.

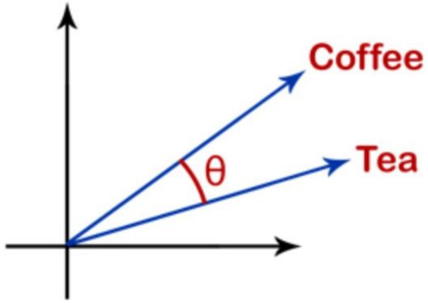
📉 Dimensionality reduction helps capture **intrinsic relations** between words and grammar.

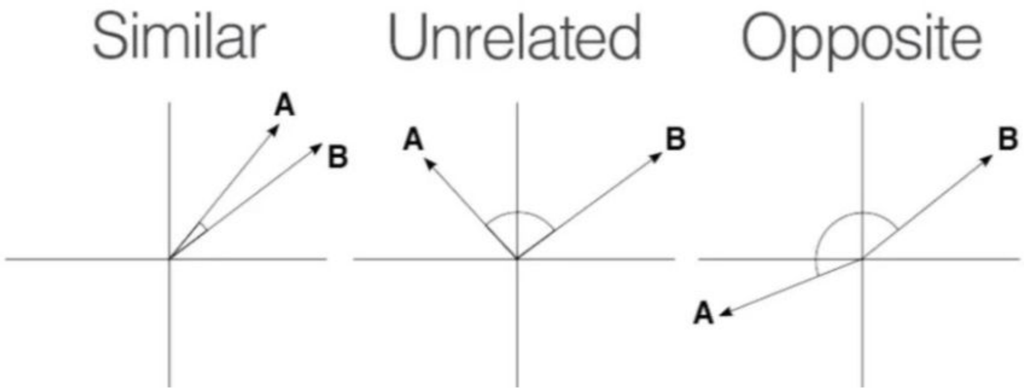
	animal	canine	pet	move	color
dog	0.81	0.8	0.89	0.68	0.56
cat	0.71	0.40	0.46	0.78	0.43
puppy	0.79	0.70	0.89	0.67	0.67
house	0.01	0.01	0.2	0.3	0.7



Similarity Measures

- Since Embeddings are vectors we can measure the cosine of the angle between the two vectors to find out the similar and dissimilar vectors





Sentence 1	Sentence 2	Cosine Similarity
The cat sits outside	The dog plays in the garden	0.2838
A man is playing guitar	A woman watches TV	-0.0327
The new movie is awesome	The new movie is so great	0.8939
Jim can run very fast	James is the fastest runner	0.6844
My goldfish is hungry	Pluto is a planet!	0.0454

BIZOM
business on the move

Our analysis of Different Embedding models

Algorithm	Precision@0.95	Precision@0.90	Precision@0.85		Recall@0.95	Recall@0.90	Recall@0.85
CLIP-VIT-L-14	1	0.8	0.2		0.95	0.93	0.88
Siglip-VIT-B-16	1	0.81	0.12		1	0.98	0.95
ClipSegrd-64-refined	1	0.98	0.82		0.84	0.77	0.6
DINO Base	1	0.99	0.88		0.55	0.5	0.4

DINO Base:
✅ **High Precision** even at lower thresholds.
⚠️ **Very low recall**, risking loss of many valid cases/scenarios
🚫 No clear threshold for **TP vs. TN separation**

ClipSegrd-64-refined:
📉 **Lower recall** compared to CLIP.
✅ **High Precision** maintained even at a low threshold of **0.85**.
🚫 No clear threshold for **TP vs. TN separation**

CLIP & SigLip
✅ Both models show high accuracy and recall at a **0.95 threshold**.
⚠️ Accuracy drops significantly at **lower thresholds**.
📉 Only **high similarity scores are reliable** for determining similarity.
✅ Clear threshold for **TP vs. TN separation**

DINO: 0.82, CLIP: 0.93

We are using an ensemble of CLIP and SIGLIP to get

- **Optimal Precision**
- **Separation of Classes (TP vs TN)**
- **Good Recall**

BIZOM
business on the move

Results from Duplicate Detection Algorithm

Different set of people are visible

The right side photo is not showing anything on the shelf while the left side photo shows the proper shop

The left side photo is the top half of the right side

The camera has been panned to take the right side photo

BIZOM
business on the move

Success Story: Scaling Image Validation Across Semi Urban India

•📸 **1 Million Images**
Captured and processed across **25 states** in semi urban India (mostly Kirana stores).

•✅ **94% Accuracy**
Achieved in classifying **valid vs. invalid** images using a **hybrid AI approach (MLLM + trained ResNet)**.

•✅ **Image Deduplication Accuracy**

•**96%** at **0.95 similarity threshold**

•**85%** at **0.90 similarity threshold**

Advantage of the Deduplication Approach

No training of Images required. Only the dump of the images is needed

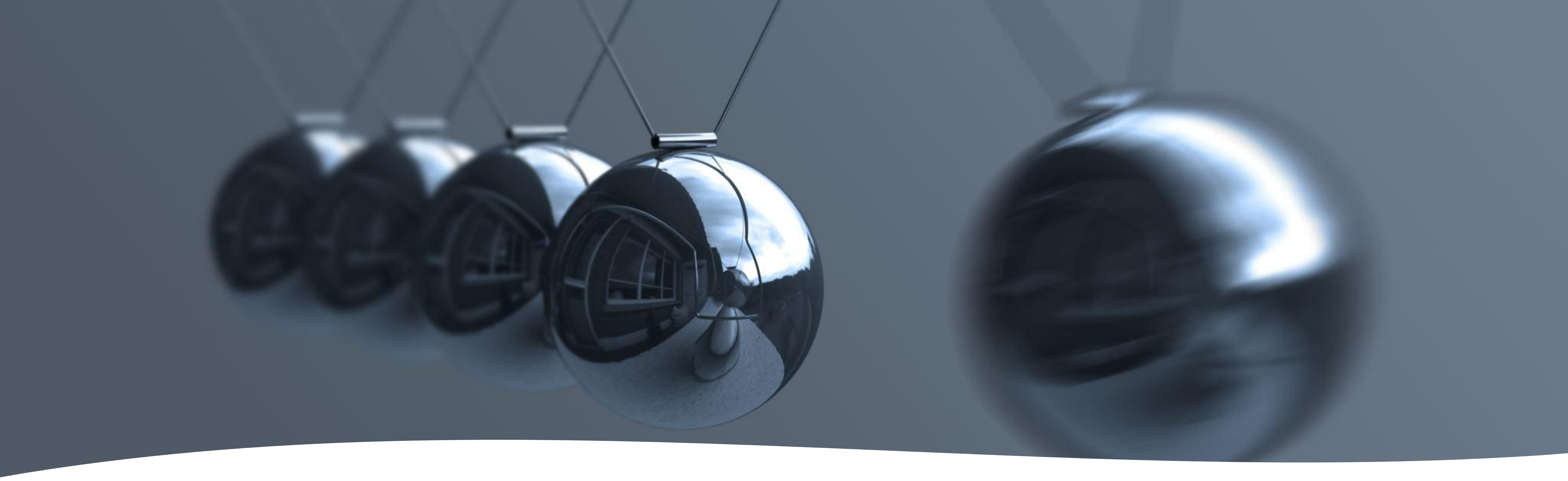
Fast execution

Works for any domain

Able to find out matches beyond the pixel comparison

Can be applied to different geographies of the world

BIZOM
business on the move



Some Extensions of the two approaches

Outlet Image Search: Finding needle in the Hay

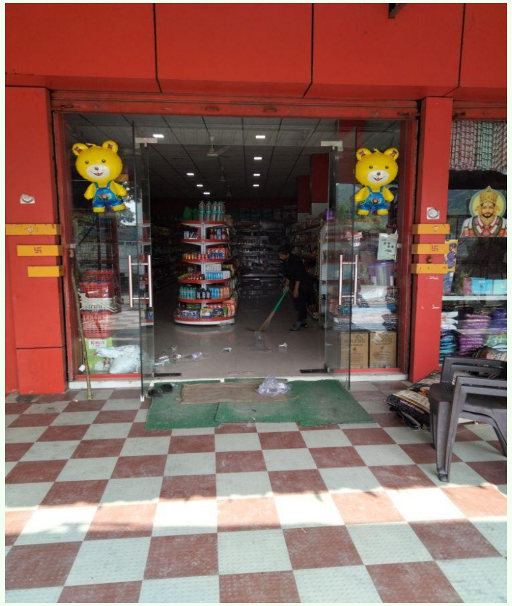
“I'm planning to launch a new product priced at $20. Which stores are likely to be ready for it?”

Approach 1 (Intuition Based)

I have the sales data I know 5$ pack sells over here let me place 20$ pack also

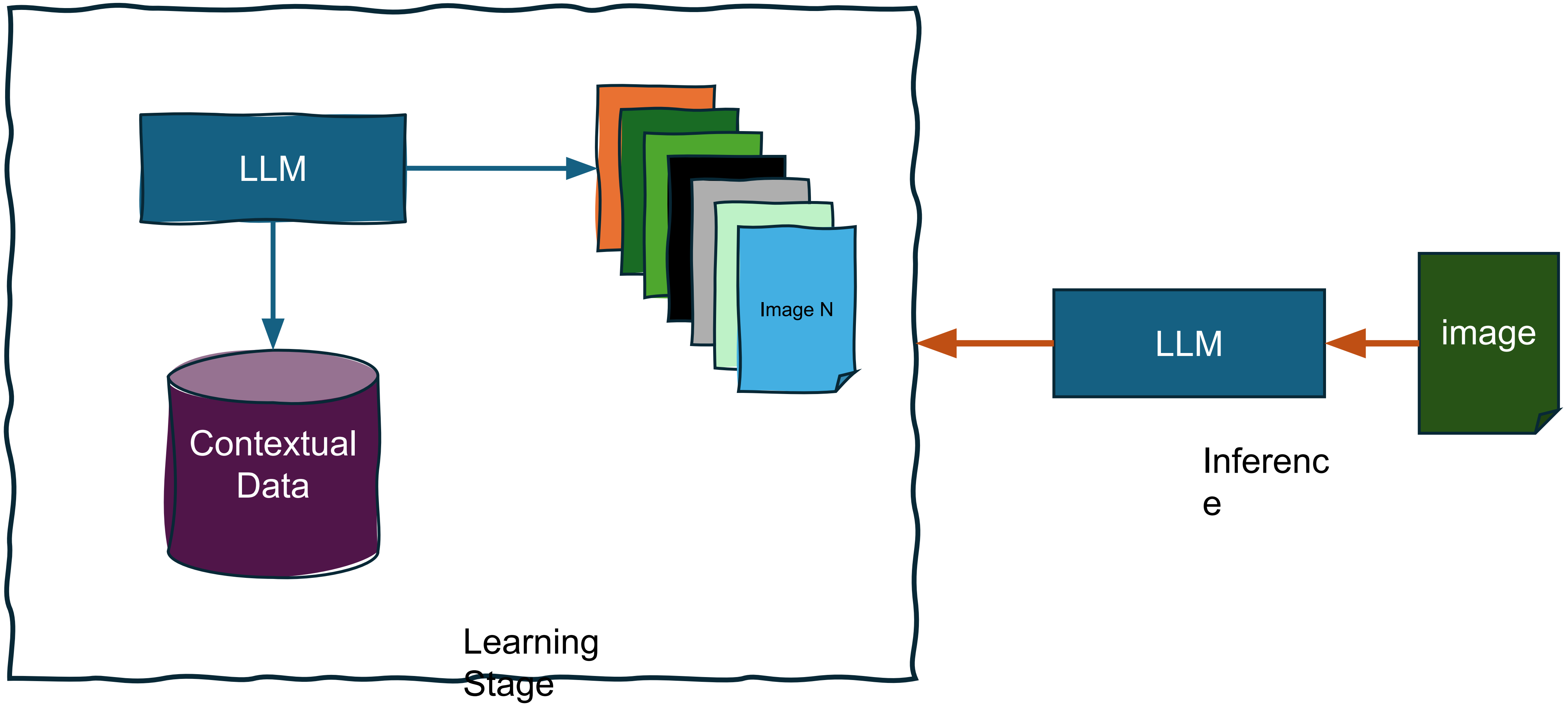
Approach 2 (Visual Inspection)

V/s

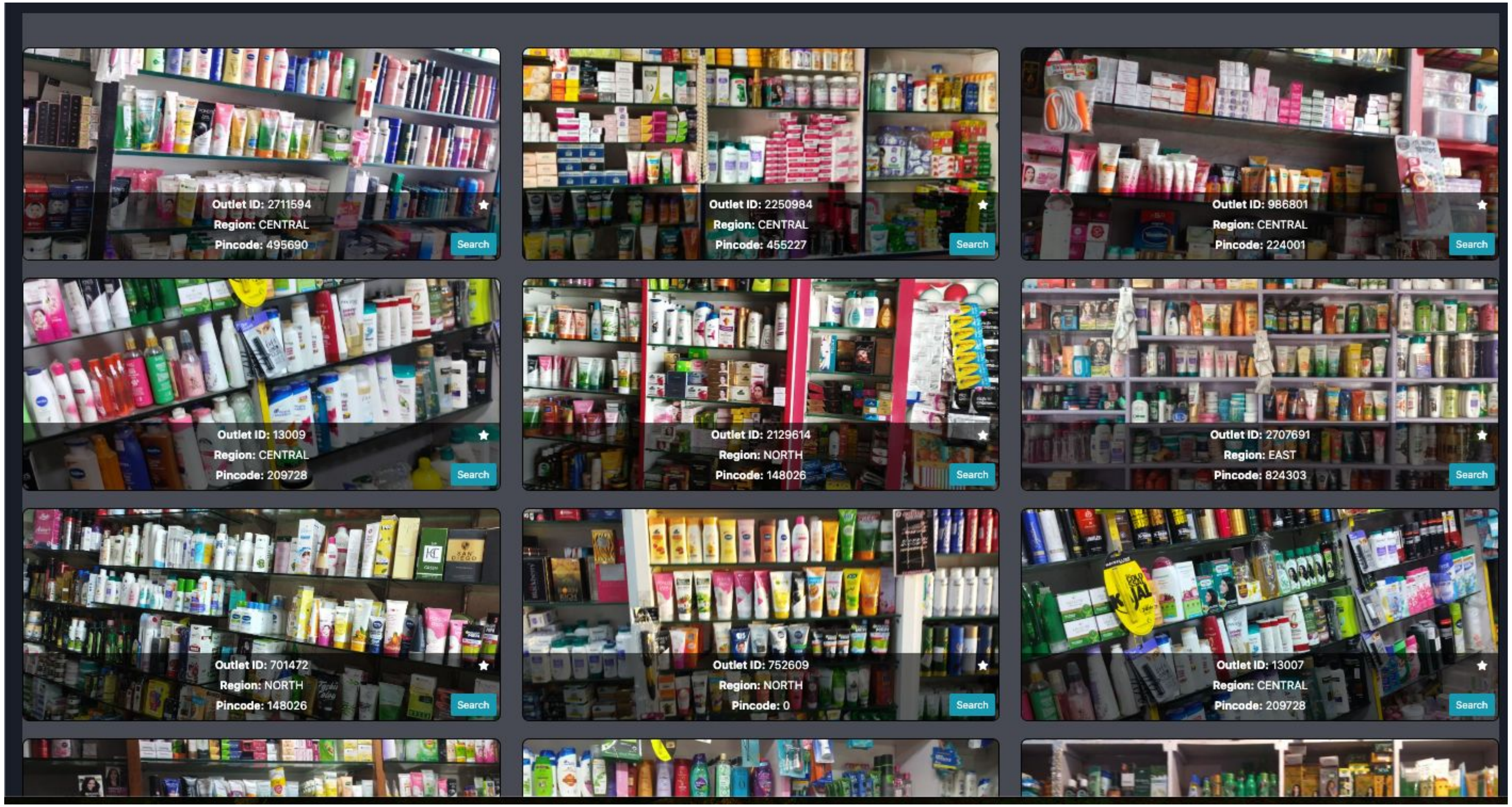


BIZOM
business on the move

Can we help in finding the right outlets?



Usecase#1: In which outlets can I place my 20$ shampoo

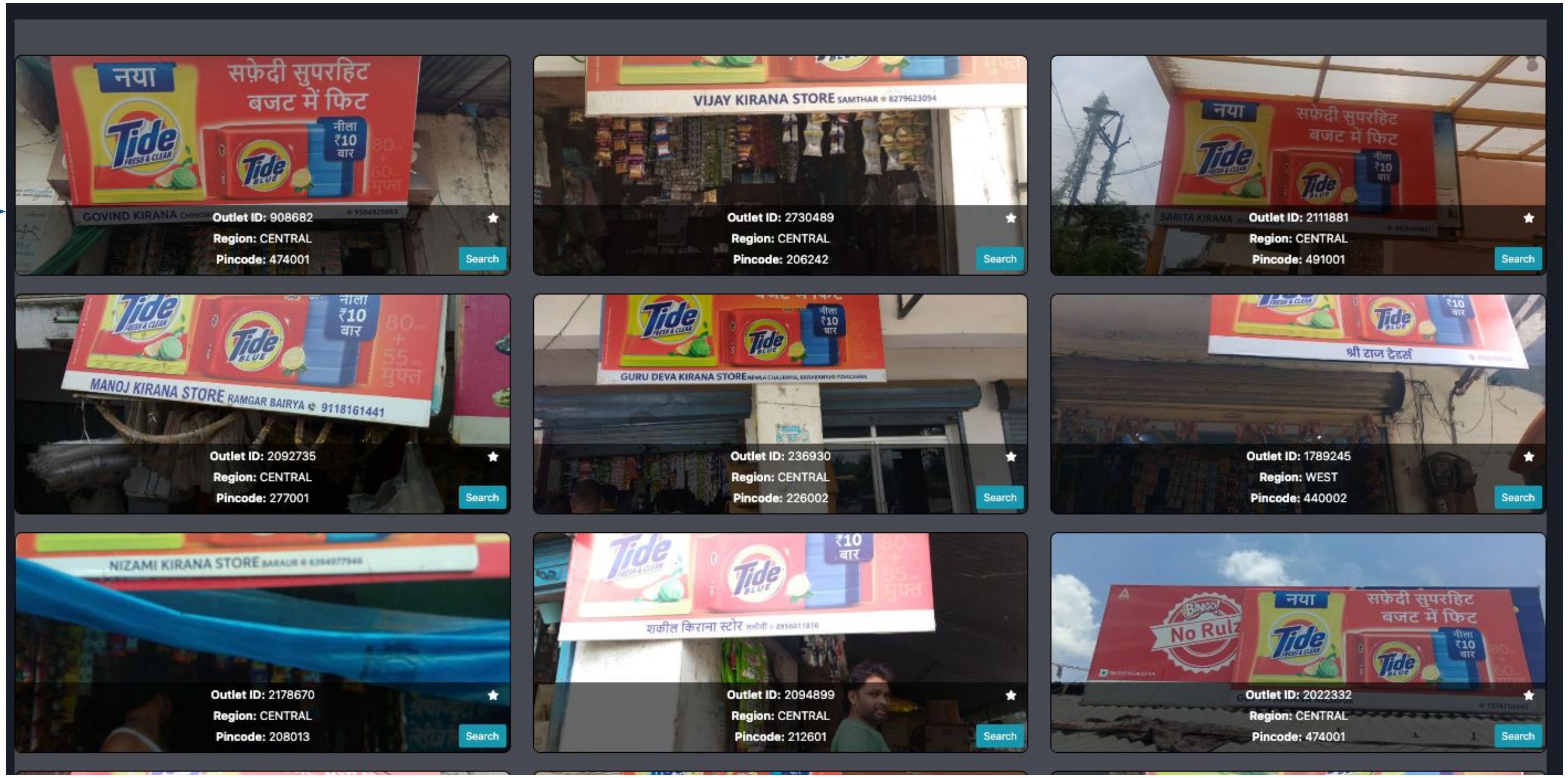


Presence of Garnier product is a checkmark that my product can be sold

BIZOM
business on the move

Usecase#2: Is my Trade spend on Advertising giving good visibility against competition

Seed Image of My Advertising

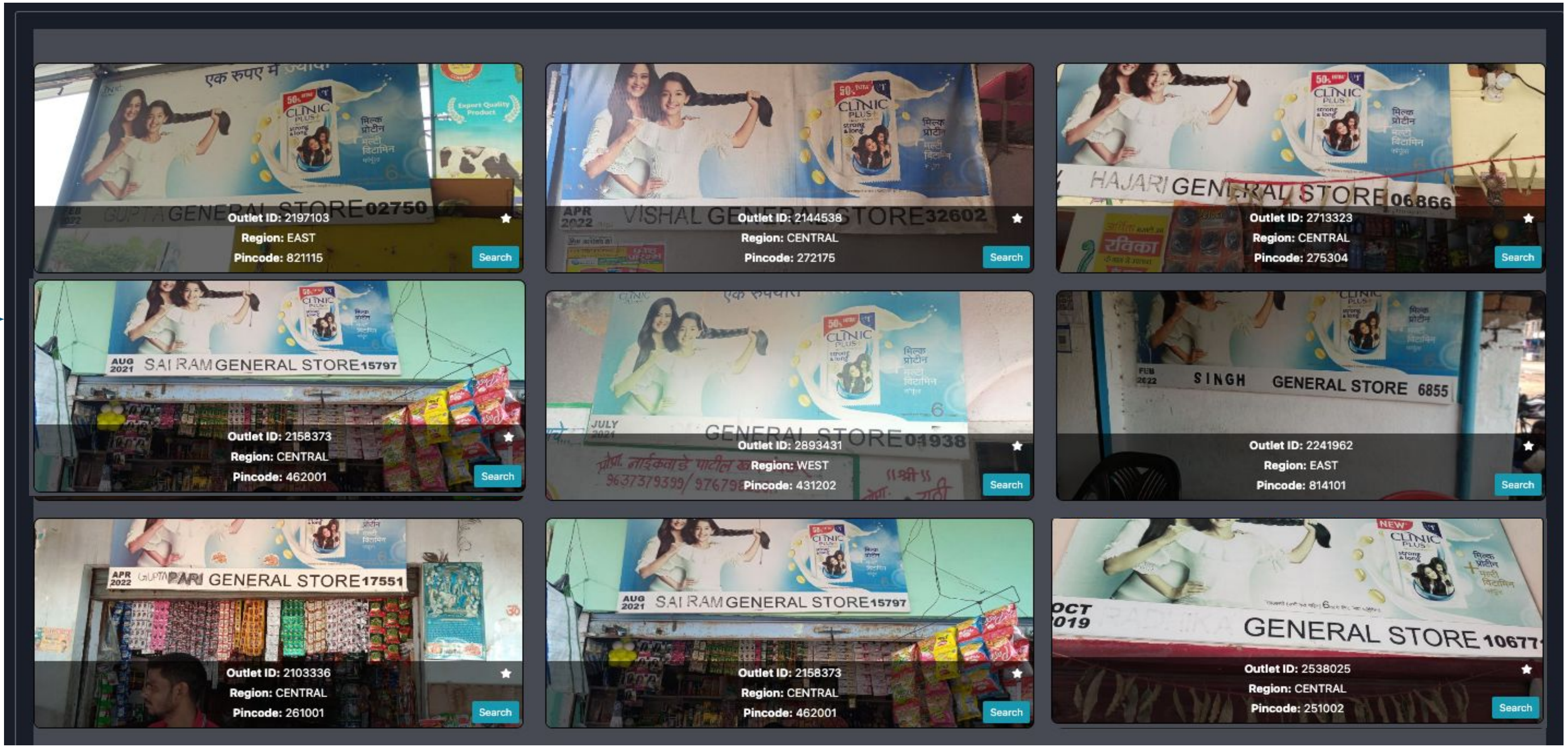


Outlets in my universe which show my advertising

BIZOM
business on the move

Usecase#1: Is my Trade spend on Advertising giving good visibility against competition

Seed Image of competition Advertising



Outlets in my universe which show my competition advertising

Data will suggest if my advertising has more visibility than the competition.
I might be adding more budget or reducing the budget for Trade promotion

BIZOM
business on the move

Thank you

rohitagarwal24

rohit@mobisy.com