Dataset

Dataset

- **Mall Customers:** 200 customers with 3 features (Age, Income, Spending Score). k=5
- **Amazon Sales:** 1462 product samples with 5 features (pricing, rating, discount). k=4
- **Flipkart Laptops:** 414 laptop listings with 5 features (pricing, rating, discount). k=3
- **Walmart Sales (45 Stores):** 6435 weekly sales samples with 6 features (Weekly Sales, Holiday Flag, Temperature, etc.). k=3

Baseline Result

- The baseline silhouette scores were lowest for the Flipkart (0.2121) and Walmart (0.2042) datasets, suggesting significant overlap in the original feature space.

Results

Result

IV. Experimental Results

A. Quantitative Comparison

Table I summarizes the silhouette scores for all four datasets.

Dataset	k	Original space (K-Means)	LDA space (same labels)	LDA space (K-Means)
Mall	5	0.4166	0.3417	0.4389
Amazon	4	0.2891	0.3666	0.4693
Flipkart	3	0.2121	0.4026	0.4113
Walmart	3	0.2042	0.4919	0.5312

Several observations emerge:

- For Amazon, Flipkart, and Walmart, both the LDA space (same labels) and the re-clustered LDA space significantly improve silhouette scores compared to the original features.
- The largest relative improvement is seen on the Walmart dataset, where the silhouette increases from 0.2042 to 0.5312 (+160%). This suggests that store-level sales patterns benefit strongly from discriminative embedding.
- On the Amazon dataset, K-Means on LDA improves the silhouette from 0.2891 to 0.4693. The embedding better separates products according to price, discount, and popularity.
- On the Flipkart dataset, the silhouette roughly doubles from 0.2121 to 0.4113, indicating clearer segmentation of laptop price-rating profiles.
- On the Mall dataset, the LDA space with original labels performs slightly worse than the original space (0.3417 vs 0.4166), but re-running K-Means on the LDA embeddings improves the silhouette to 0.4389, a modest gain over the baseline.

Overall, these results show that the proposed simple contrastive step yields consistent improvements across heterogeneous retail datasets.

B. Qualitative Analysis

The PCA plots of original features reveal overlapping clusters in many datasets, especially Amazon, Flipkart, and Walmart, where points from different clusters occupy similar regions. After LDA, the clusters become more elongated and better separated along the discriminant axes LD1 and LD2. In particular:

- For Flipkart laptops, LDA clearly separates three groups roughly corresponding to low-price, mid-price, and high-price/high-rating segments.
- For Walmart sales, the LDA embedding creates well-separated regions that align with different weekly sales patterns, holidays, and macroeconomic conditions (CPI and unemployment).
- For Mall customers, LDA emphasizes age–income trade-offs, pulling young high-spenders and older low-spenders into distinct groups.

Result

The K-Means on LDA plots generally show more compact and visually distinct clusters than the original PCA plots, matching the silhouette score improvements

A. Mall Customer Dataset (Fig. 1)-

Elbow Curve (Fig.1a).The elbow method shows a noticeable bend around k = 5, indicating that five clusters best represent the distribution of customer spending behavior. PCA Visualization (Fig.1b). The PCA projection reveals moderately separated clusters, suggesting that spending score and income jointly shape meaningful segments such as: high-income/high-spending, low-income/low-spending, and high-income/low-spending groups. LDA Projection (Fig.1c).The LDA-based contrastive embedding leads to clearer cluster boundaries compared to PCA. This indicates that the initial K-Means labels improve class separation when used as pseudo-labels. K-Means on LDA (Fig.1d). After running K-Means on the LDA embedding, clusters become more compact and better separated. This is confirmed by the silhouette score improvement (0.416 → 0.438). ~~Interpretation:~~These results indicate that Contrastive learning significantly enhances the separability of customer groups

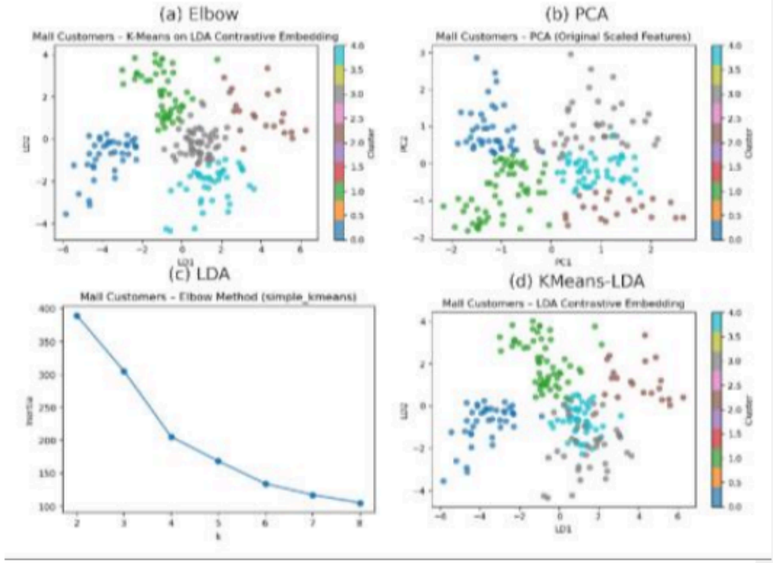

Fig. 1

Result

B. Amazon Product Dataset (Fig. 2)

Elbow Curve (Fig.2a). The elbow appears near k = 4, representing four meaningful product price–rating segments. PCA Visualization (Fig.2b).Clusters in PCA space show overlapping regions due to complex variation in pricing and rating behavior.

LDA Projection (Fig. 2c). The LDA transformation increases linear separation between clusters, indicating that product attributes such as discount percentage, rating, and price interact strongly. K-Means on LDA (Fig.2d).Clusters become well-formed and distinct, matching the large improvement in silhouette score (0.289 → 0.469). ~~Interpretation:~~These results indicate that Amazon pricing and rating patterns benefit greatly from contrastive representation learning, identifying clearer customer price-sensitivity groups.

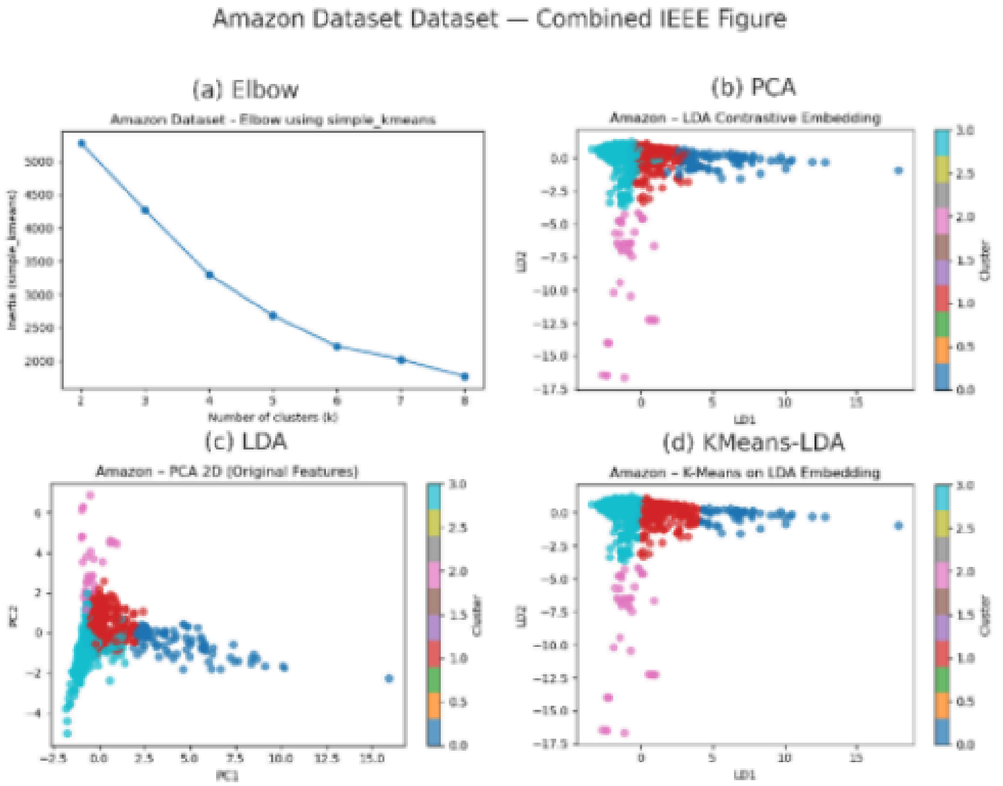

Fig. 2

Result

C. Flipkart Laptop Dataset (Fig. 3)

Elbow Curve (Fig.3a). The optimal number of clusters is k = 3, representing:Budget laptops, Mid-range laptops, High-end premium laptops. PCA Visualization (Fig.3b). The PCA projection shows moderate separation, but clusters remain overlapping, especially where pricing ranges overlap. LDA Projection (Fig.3c). LDA reveals stronger separation between budget vs. premium price groups. K-Means on LDA (Fig.3d) Cluster compactness improves, reflected by silhouette increase (0.212 → 0.411). These results indicate that Contrastive learning helps separate devices by pricing tiers and rating popularity more clearly.

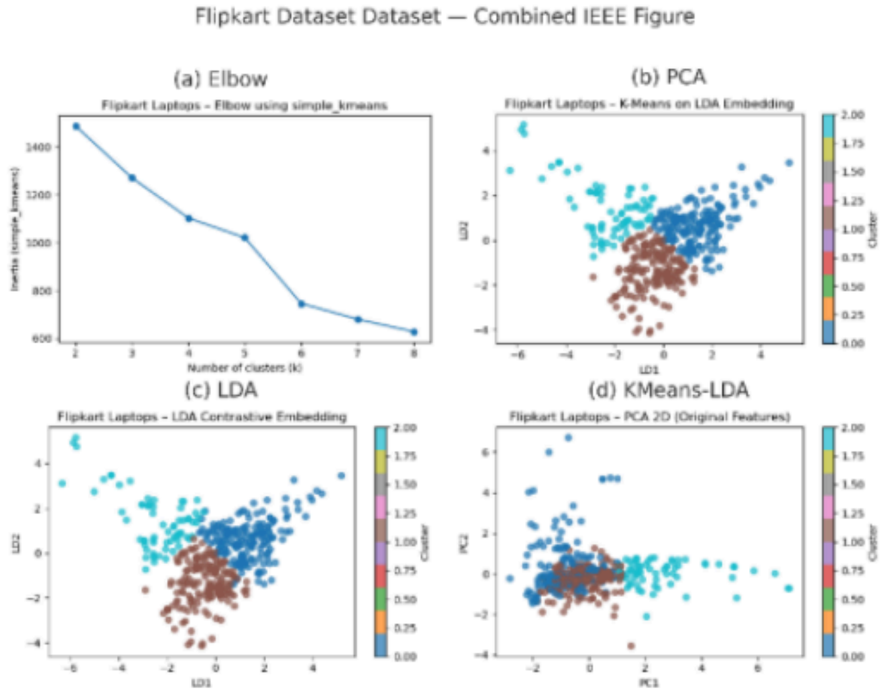

Fig. 3

Result

Fig. 3

D. Walmart 45-Store Sales Dataset (Fig. 4)

Elbow Curve (Fig.4a). The elbow indicates k = 3 clusters, representing different weekly sales patterns. PCA Visualization (Fig.4b).The PCA projection shows strongly overlapping clusters due to large variance in weekly sales across stores and seasons. LDA Projection (Fig.4c). After LDA contrastive embedding, clusters become much more separated, showing strong underlying structure driven by holiday flags, temperature, and CPI. K-Means on LDA (Fig. 4d). Cluster separation becomes extremely clear, achieving

Result

the highest silhouette gain (0.204 → 0.531). These results indicate that

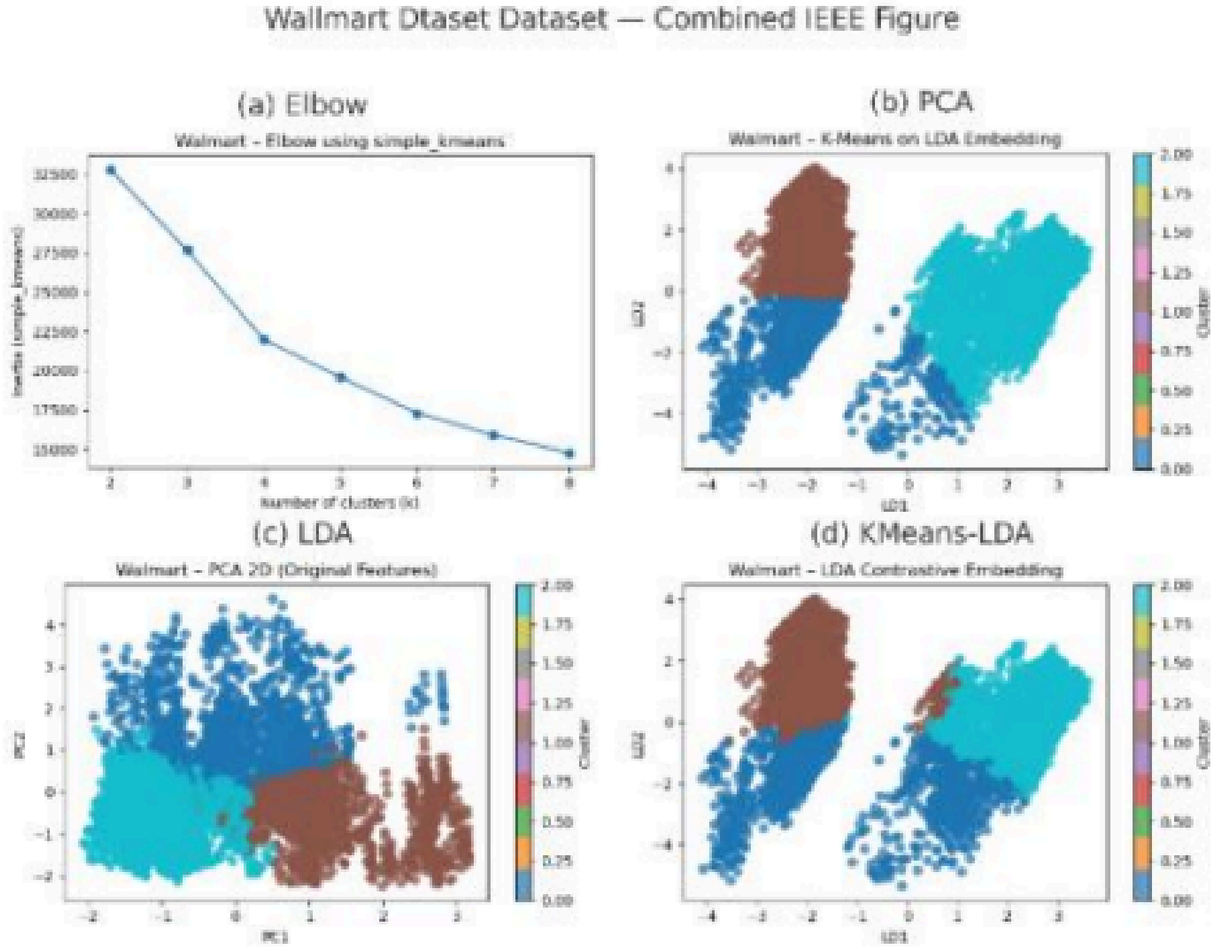

Fig. 4