A Lightweight Image Understanding System for Classification, Attribute Prediction, and Retrieval

FIRST SEMESTER 2025-2026

UNDER THE SUPERVISION OF

Dr. Pratik Narang

Department of CSIS

TEAM MEMBERS

Rajiv Reddy Kasarla	2024H1030080P
Eswar Narayana	2024H1120194P
Shreedhar Soni	2024H1120179P

1. Overview

This milestone focused on developing a lightweight image-understanding system that can

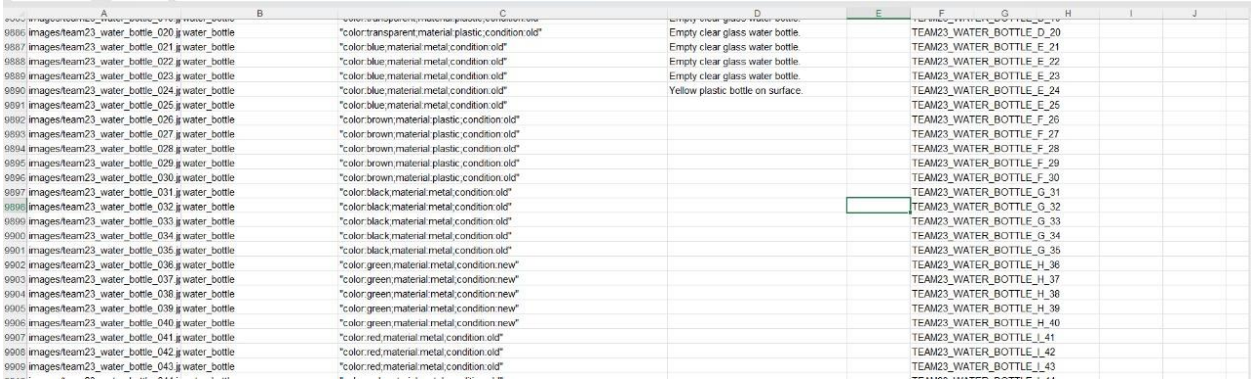
1. classify an image into one of several object classes,
2. predict its visual attributes such as color, material, and condition, and
3. perform text → image retrieval.

Two **Vision Transformer (ViT)** variants were fine-tuned on two datasets our **own Dataset 1** and the **pooled Dataset 2** to compare the impact of dataset scale and diversity on performance and generalization.

2. Datasets

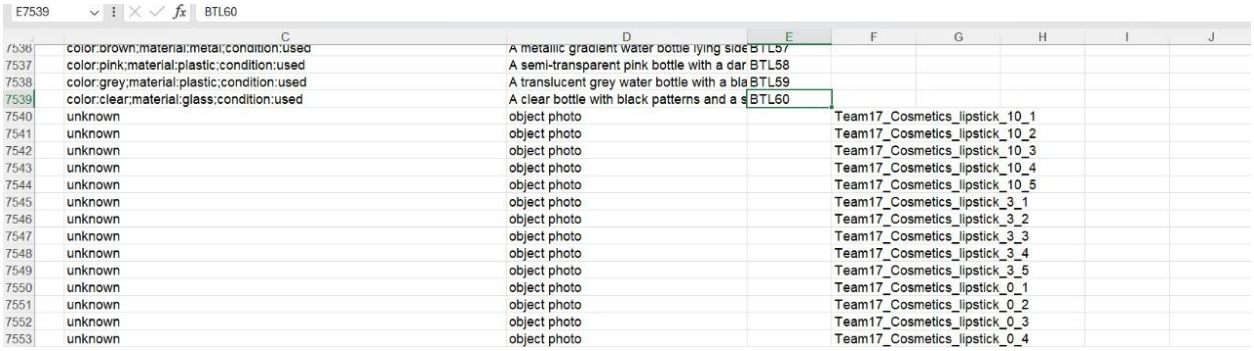
Property	**Dataset 1 (Own)**	**Dataset 2 (Pooled)**
No. of Classes	10	197
Colors / Materials / Conditions	15 / 9 / 4	33 / 39 / 14

Dataset 1 was small but clean and balanced: each class had ~60 images with clear attribute annotations.
Dataset 2 was large but noisy and inconsistent many images had missing or invalid labels ("unknown"), incomplete captions, and redundant classes such as *water_bottle_1 – 10*, which greatly affected training quality.



	A	B	C	D	E	F	G	H	I	J
9886	images/team23_water_bottle_020.jpg	water_bottle	"color:transparent;material:plastic;condition:old"	Empty clear glass water bottle.		TEAM23_WATER_BOTTLE_D_20				
9887	images/team23_water_bottle_021.jpg	water_bottle	"color:blue;material:metal;condition:old"	Empty clear glass water bottle.		TEAM23_WATER_BOTTLE_E_21				
9888	images/team23_water_bottle_022.jpg	water_bottle	"color:blue;material:metal;condition:old"	Empty clear glass water bottle.		TEAM23_WATER_BOTTLE_E_22				
9889	images/team23_water_bottle_023.jpg	water_bottle	"color:blue;material:metal;condition:old"	Empty clear glass water bottle.		TEAM23_WATER_BOTTLE_E_23				
9890	images/team23_water_bottle_024.jpg	water_bottle	"color:blue;material:metal;condition:old"	Yellow plastic bottle on surface.		TEAM23_WATER_BOTTLE_E_24				
9891	images/team23_water_bottle_025.jpg	water_bottle	"color:blue;material:metal;condition:old"			TEAM23_WATER_BOTTLE_E_25				
9892	images/team23_water_bottle_026.jpg	water_bottle	"color:brown;material:plastic;condition:old"			TEAM23_WATER_BOTTLE_F_26				
9893	images/team23_water_bottle_027.jpg	water_bottle	"color:brown;material:plastic;condition:old"			TEAM23_WATER_BOTTLE_F_27				
9894	images/team23_water_bottle_028.jpg	water_bottle	"color:brown;material:plastic;condition:old"			TEAM23_WATER_BOTTLE_F_28				
9895	images/team23_water_bottle_029.jpg	water_bottle	"color:brown;material:plastic;condition:old"			TEAM23_WATER_BOTTLE_F_29				
9896	images/team23_water_bottle_030.jpg	water_bottle	"color:brown;material:plastic;condition:old"			TEAM23_WATER_BOTTLE_F_30				
9897	images/team23_water_bottle_031.jpg	water_bottle	"color:black;material:metal;condition:old"			TEAM23_WATER_BOTTLE_G_31				
9898	images/team23_water_bottle_032.jpg	water_bottle	"color:black;material:metal;condition:old"			TEAM23_WATER_BOTTLE_G_32				
9899	images/team23_water_bottle_033.jpg	water_bottle	"color:black;material:metal;condition:old"			TEAM23_WATER_BOTTLE_G_33				
9900	images/team23_water_bottle_034.jpg	water_bottle	"color:black;material:metal;condition:old"			TEAM23_WATER_BOTTLE_G_34				
9901	images/team23_water_bottle_035.jpg	water_bottle	"color:black;material:metal;condition:old"			TEAM23_WATER_BOTTLE_G_35				
9902	images/team23_water_bottle_036.jpg	water_bottle	"color:green;material:metal;condition:new"			TEAM23_WATER_BOTTLE_H_36				
9903	images/team23_water_bottle_037.jpg	water_bottle	"color:green;material:metal;condition:new"			TEAM23_WATER_BOTTLE_H_37				
9904	images/team23_water_bottle_038.jpg	water_bottle	"color:green;material:metal;condition:new"			TEAM23_WATER_BOTTLE_H_38				
9905	images/team23_water_bottle_039.jpg	water_bottle	"color:green;material:metal;condition:new"			TEAM23_WATER_BOTTLE_H_39				
9906	images/team23_water_bottle_040.jpg	water_bottle	"color:green;material:metal;condition:new"			TEAM23_WATER_BOTTLE_H_40				
9907	images/team23_water_bottle_041.jpg	water_bottle	"color:red;material:metal;condition:old"			TEAM23_WATER_BOTTLE_I_41				
9908	images/team23_water_bottle_042.jpg	water_bottle	"color:red;material:metal;condition:old"			TEAM23_WATER_BOTTLE_I_42				
9909	images/team23_water_bottle_043.jpg	water_bottle	"color:red;material:metal;condition:old"			TEAM23_WATER_BOTTLE_I_43				

Blank captions



	C	D	E	F	G	H	I	J
7536	color:brown;material:metal;condition:used	A metallic gradient water bottle lying side	BTL57					
7537	color:pink;material:plastic;condition:used	A semi-transparent pink bottle with a dar	BTL58					
7538	color:grey;material:plastic;condition:used	A translucent grey water bottle with a bla	BTL59					
7539	color:clear;material:glass;condition:used	A clear bottle with black patterns and a s	BTL60					
7540	unknown	object photo		Team17_Cosmetics_lipstick_10_1				
7541	unknown	object photo		Team17_Cosmetics_lipstick_10_2				
7542	unknown	object photo		Team17_Cosmetics_lipstick_10_3				
7543	unknown	object photo		Team17_Cosmetics_lipstick_10_4				
7544	unknown	object photo		Team17_Cosmetics_lipstick_10_5				
7545	unknown	object photo		Team17_Cosmetics_lipstick_3_1				
7546	unknown	object photo		Team17_Cosmetics_lipstick_3_2				
7547	unknown	object photo		Team17_Cosmetics_lipstick_3_3				
7548	unknown	object photo		Team17_Cosmetics_lipstick_3_4				
7549	unknown	object photo		Team17_Cosmetics_lipstick_3_5				
7550	unknown	object photo		Team17_Cosmetics_lipstick_0_1				
7551	unknown	object photo		Team17_Cosmetics_lipstick_0_2				
7552	unknown	object photo		Team17_Cosmetics_lipstick_0_3				
7553	unknown	object photo		Team17_Cosmetics_lipstick_0_4				

Wrong description /caption and unknown attributes

10810	images/team11_water_bottle_004.	water_bottle	color:transparent:material:plastic:condition:new	water_bottle image
10811	images/team11_water_bottle_005.	water_bottle	color:transparent:material:plastic:condition:new	water_bottle image
10812	images/team11_water_bottle_006.	water_bottle	color:transparent:material:plastic:condition:new	water_bottle image
10813	images/team11_water_bottle_007.	water_bottle	color:transparent:material:plastic:condition:new	water_bottle image
10814	images/team11_water_bottle_008.	water_bottle	color:pink;material:metal;condition:new	water_bottle image
10815	images/team11_water_bottle_1_0(	water_bottle_1	color:pink;material:metal;condition:new	water_bottle_1 image
10816	images/team11_water_bottle_1_0(	water_bottle_1	color:pink;material:metal;condition:new	water_bottle_1 image
10817	images/team11_water_bottle_1_0(	water_bottle_1	color:pink;material:metal;condition:new	water_bottle_1 image
10818	images/team11_water_bottle_1_0(	water_bottle_1	color:yellow;material:metal;condition:new	water_bottle_1 image
10819	images/team11_water_bottle_1_0(	water_bottle_1	color:pink;material:metal;condition:new	water_bottle_1 image
10820	images/team11_water_bottle_10_(	water_bottle_10	color:red;material:metal;condition:new	water_bottle_10 image
10821	images/team11_water_bottle_10_(	water_bottle_10	color:transparent:material:plastic:condition:new	water_bottle_10 image
10822	images/team11_water_bottle_10_(	water_bottle_10	color:transparent:material:plastic:condition:new	water_bottle_10 image
10823	images/team11_water_bottle_10_(	water_bottle_10	color:transparent:material:plastic:condition:new	water_bottle_10 image
10824	images/team11_water_bottle_10_(	water_bottle_10	color:gray:material:metal:condition:new	water_bottle_10 image
10825	images/team11_water_bottle_11_(	water_bottle_11	color:gray:material:metal:condition:new	water_bottle_11 image
10826	images/team11_water_bottle_11_(	water_bottle_11	color:gray:material:metal:condition:new	water_bottle_11 image
10827	images/team11_water_bottle_2_0(	water_bottle_2	color:gray:material:metal:condition:new	water_bottle_2 image

Wrong class names

3. Models and Training Setup

Two lightweight ViT architectures were selected:

Model	Type	Parameters	Pretraining	Fine-Tuning Phases
DeiT-Tiny	Pure ViT	≈ 5 M	ImageNet	2 phases (Head training + Full fine-tune)
MobileViT-XXS	Hybrid CNN + Transformer	≈ 1.2 M	ImageNet	2 phases (Head training + Full fine-tune)

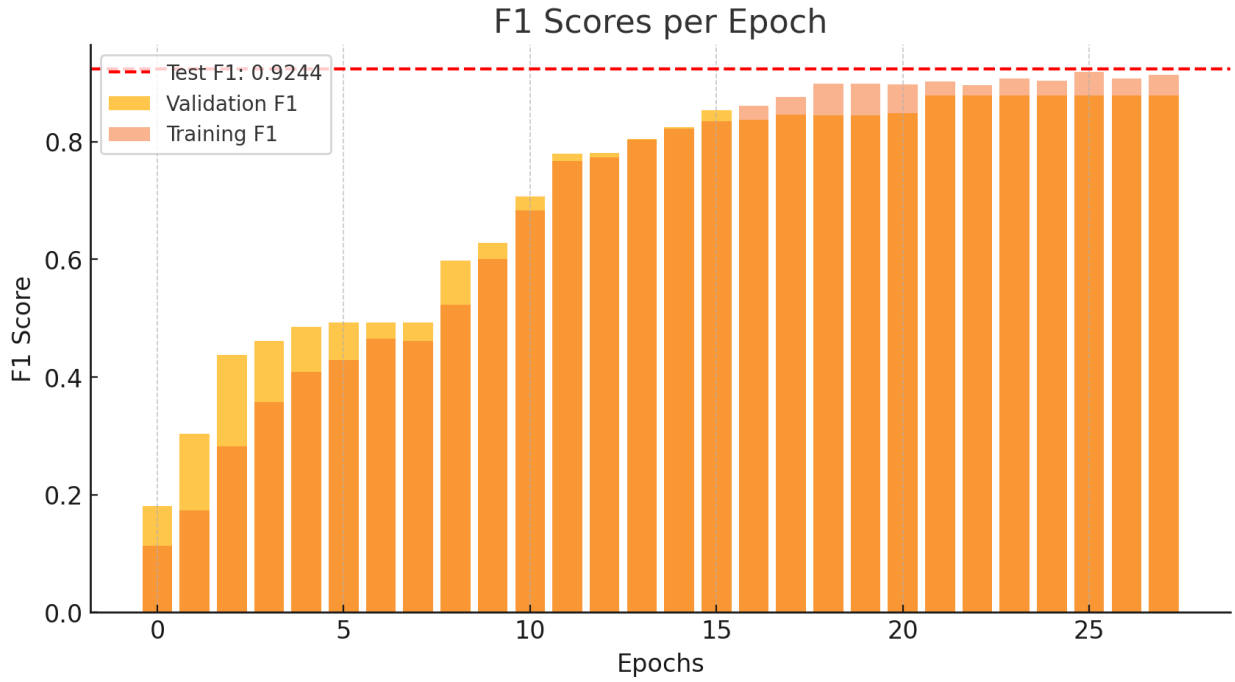
Both used identical settings:

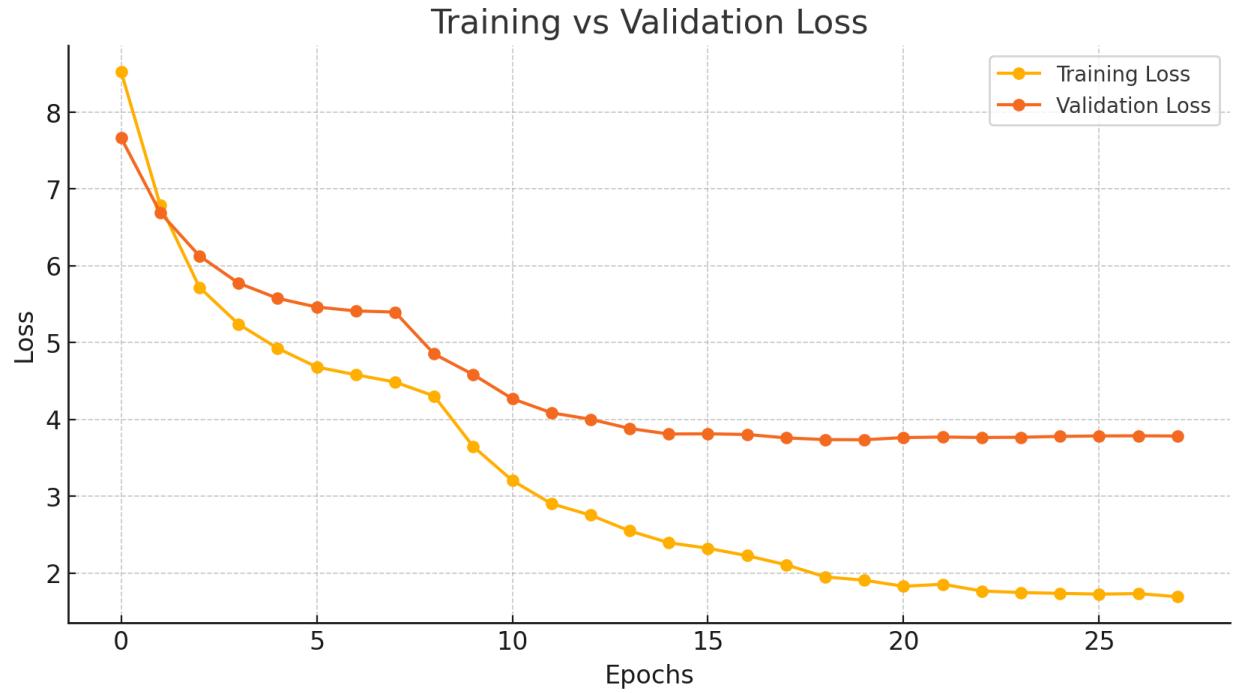
- Optimizer = AdamW, LR = 3e-4
- Epochs: 8 (phase 1) + 20 (phase 2)
- Loss: Cross-Entropy (+ attribute heads for color/material/condition)
- Input size = 224×224

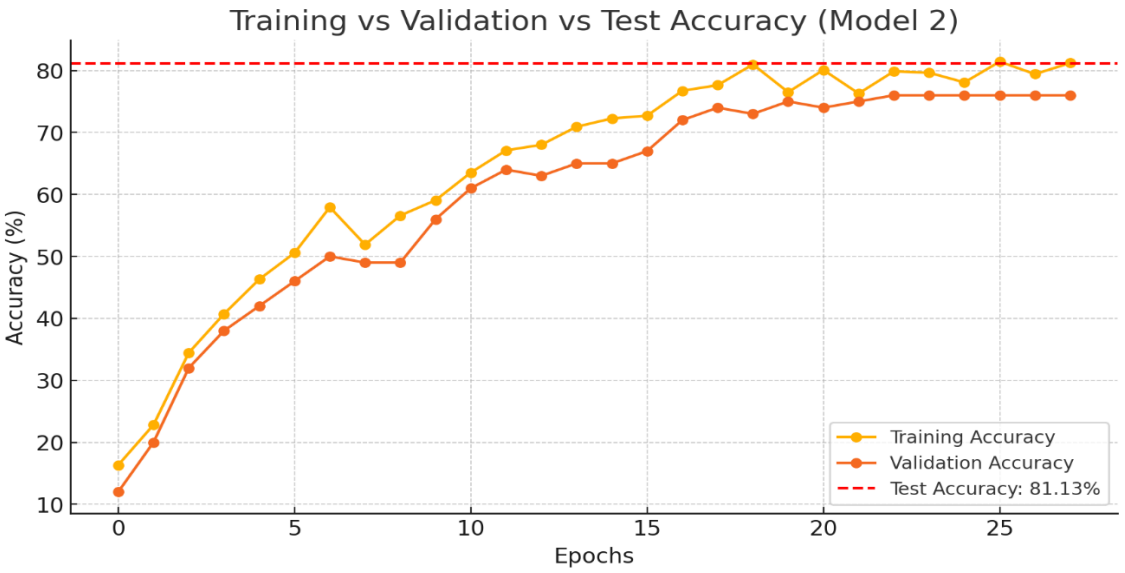
- Device = CPU / CUDA (for reproducibility)

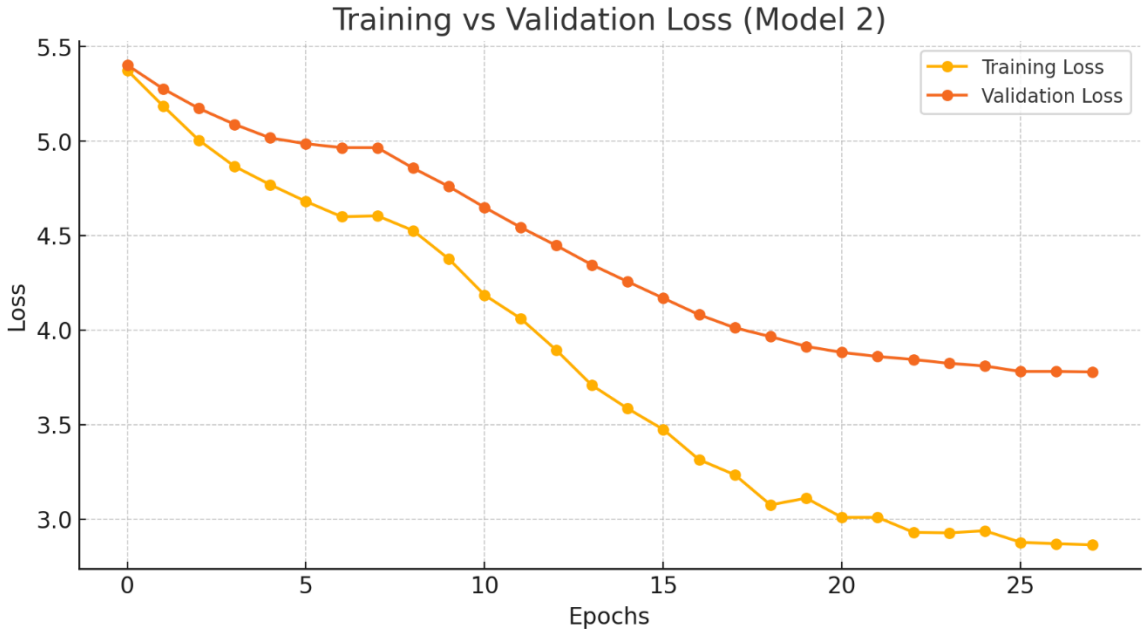
4. Results

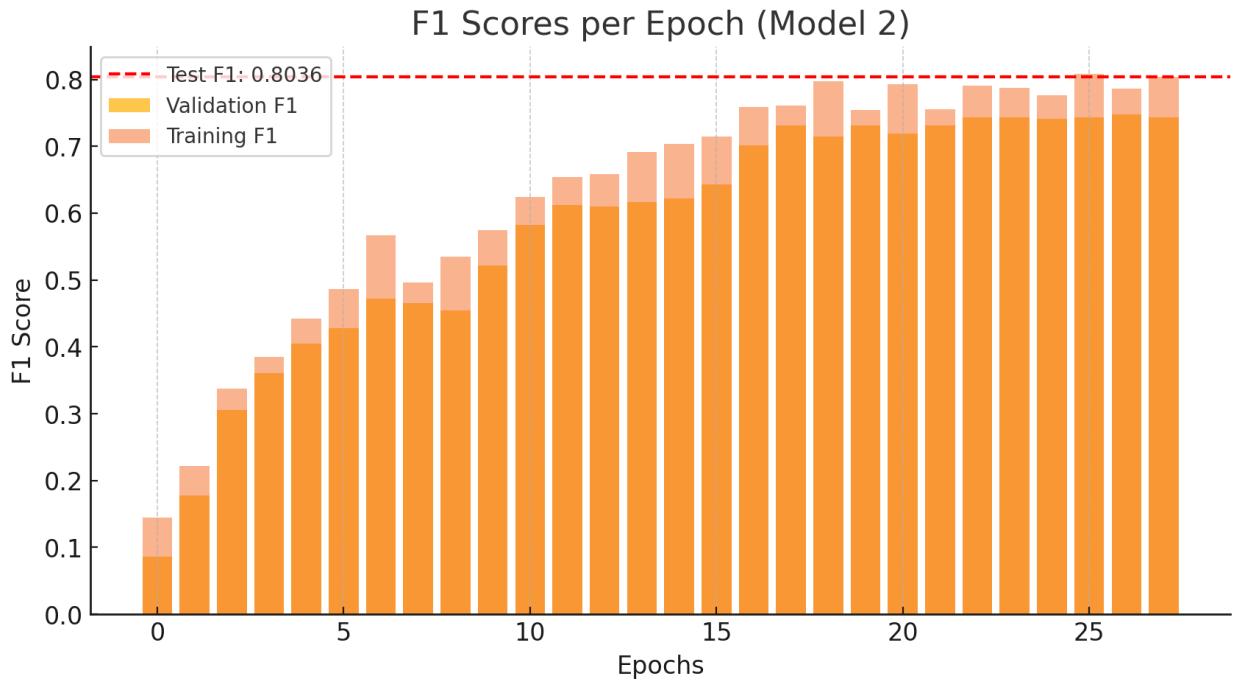
4.1 Dataset 1 (Own Data)






Training vs Validation vs Test Accuracy (Model 2)
Accuracy (%)
Epochs
Training Accuracy
Validation Accuracy
Test Accuracy: 81.13%

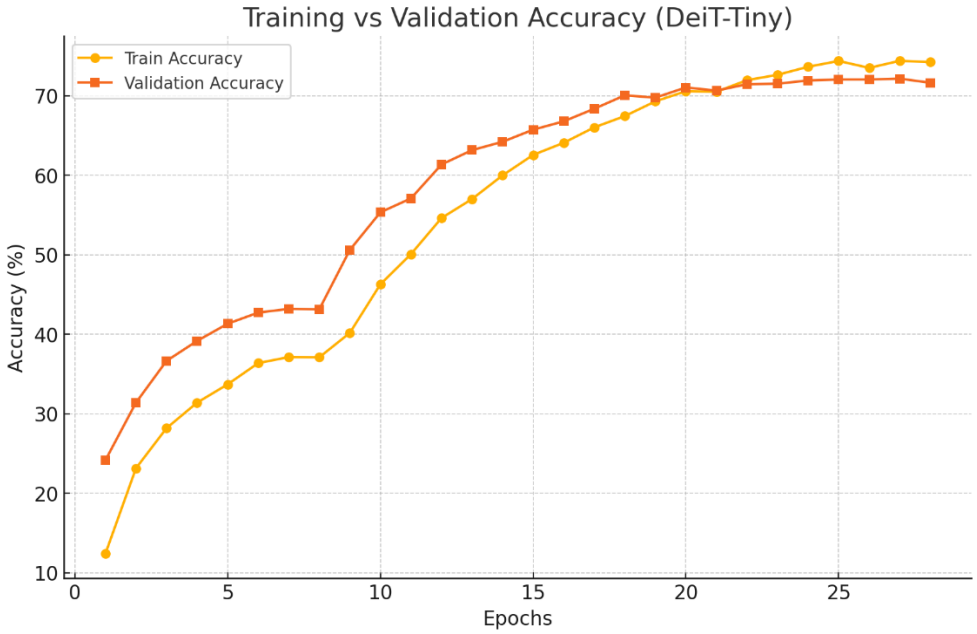

Training vs Validation Loss (Model 2)
Loss
Epochs
Training Loss
Validation Loss

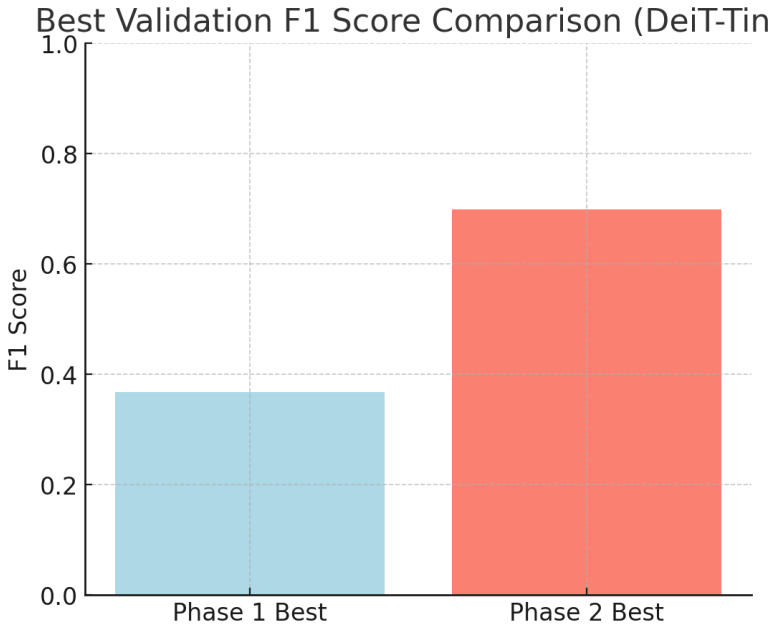


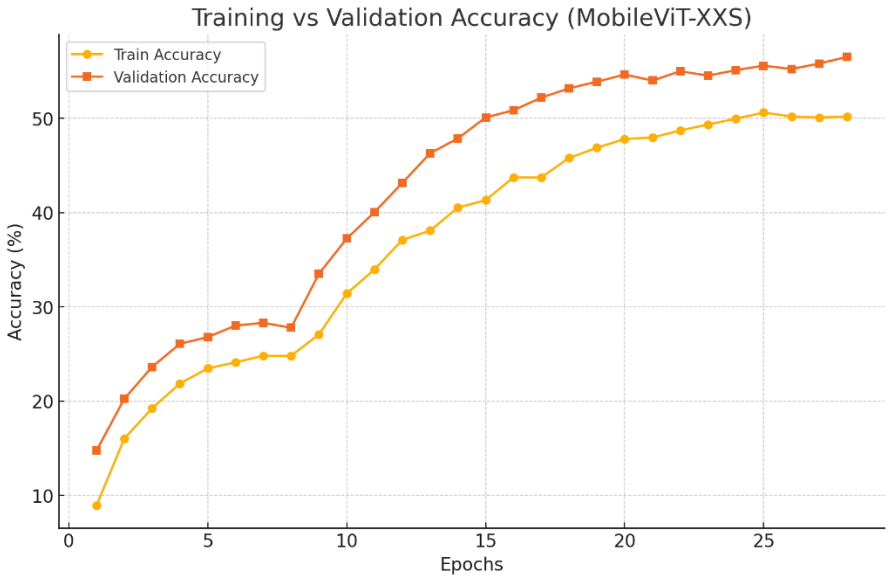
Model	Best Val Acc	Best Val F1	Test Acc	Test F1
DeiT-Tiny	92.45 %	0.924	92.45 %	0.924
MobileViT-XXS	81.13 %	0.804	81.13 %	0.804

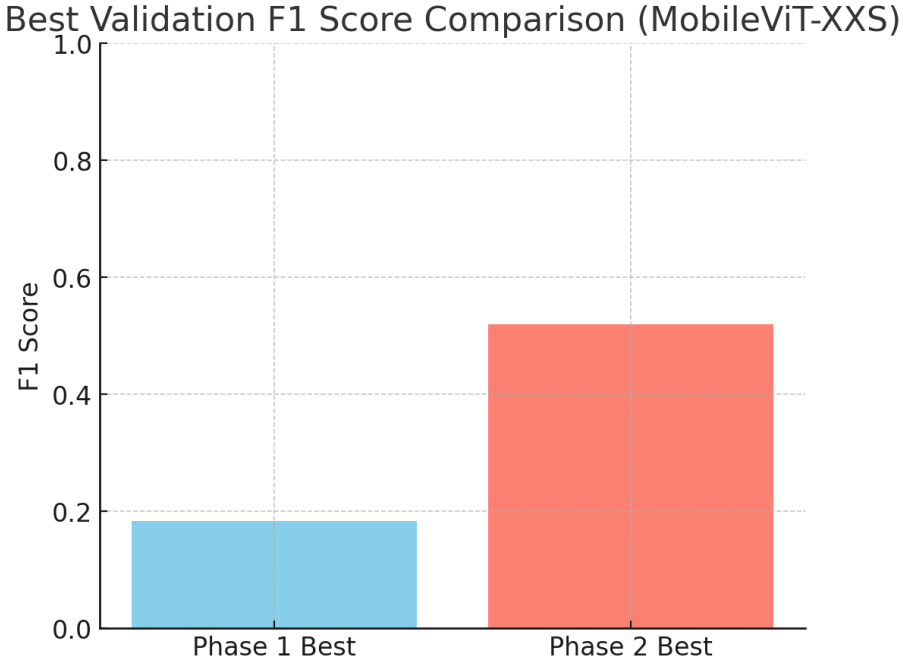
Both models converged quickly within 5 epochs of full fine-tuning. DeiT-Tiny achieved near-perfect classification, reflecting the dataset's balanced and limited scope. MobileViT-XXS also generalized well but lagged ~10 % behind in F1 due to its smaller capacity and simplified transformer blocks.

4.2 Dataset 2 (Pooled Data)


Training vs Validation Accuracy (DeiT-Tiny)
Train Accuracy
Validation Accuracy
Accuracy (%)
Epochs
10
20
30
40
50
60
70
0
5
10
15
20
25


Best Validation F1 Score Comparison (DeiT-Tiny)
F1 Score
0.0
0.2
0.4
0.6
0.8
1.0
Phase 1 Best
Phase 2 Best





Model	**Best Val Acc**	**Best Val F1**
DeiT-Tiny	72.15 %	0.6989
MobileViT-XXS	56.51 %	0.5192

Despite the larger dataset, overall scores dropped drastically compared to Dataset 1. Learning curves showed **slower convergence, higher loss variance, and weaker F1-improvement** across epochs.

5. Analysis

5.1 Model vs Model

Aspect	**DeiT-Tiny**	**MobileViT-XXS**
Learning Speed	Faster convergence; smoother loss curve	Slightly slower; more unstable early epochs
Accuracy Trend	Higher across both datasets	Lower but more stable on small dataset
Capacity vs Overfitting	More capacity → minor overfit risk on small data	Underfits on large noisy data
Attribute Prediction	Better color/material alignment	Weaker attribute head performance

Observation: DeiT-Tiny benefited from its richer token mixing and positional embeddings, achieving stronger representations.
MobileViT's local CNN inductive bias helped on simple backgrounds but struggled on diverse or mislabeled pooled data.

5.2 Dataset vs Dataset

Finding	**Explanation**
Performance ↓ in pooled data	Excess noise, mislabeled and duplicate classes caused gradient confusion
Class imbalance and redundancy	Many rare classes (< 5 samples), and redundant categories (e.g., *water_bottle_1 – 10*) inflated the softmax space to 197 labels
Attribute noise	Frequent "unknown" entries for color/material → no signal for auxiliary heads
Poor captions	Several captions only stated the object name (e.g., "photo of a bottle") → weak text–image alignment for retrieval

Black caption images	Corrupted or placeholder entries disrupted mini-batch gradients
Generalization	Diverse data should improve robustness, but only if labels are consistent; here, label entropy outweighed benefit

5.3 Why Dataset 2 Underperformed

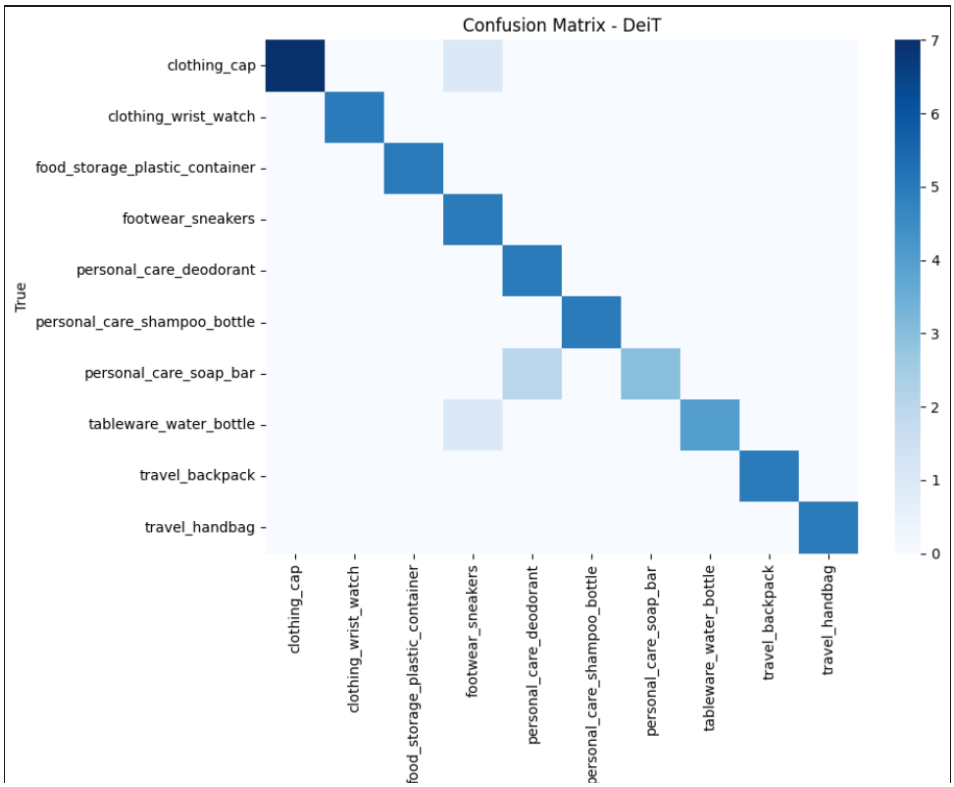
1. **Unknown attributes** → attribute heads received uninformative gradients.
2. **Caption deficiency** → retrieval head failed to map text queries (e.g., "blue plastic jug") to relevant visual concepts.
3. **Structural class flaws** → redundant classes like water_bottle_1…10 made class boundaries arbitrary.
4. **Label noise and black images** → reduced effective dataset size and increased training instability.
5. **Imbalanced attribute distribution** → certain colors/materials never appeared in validation or test splits.

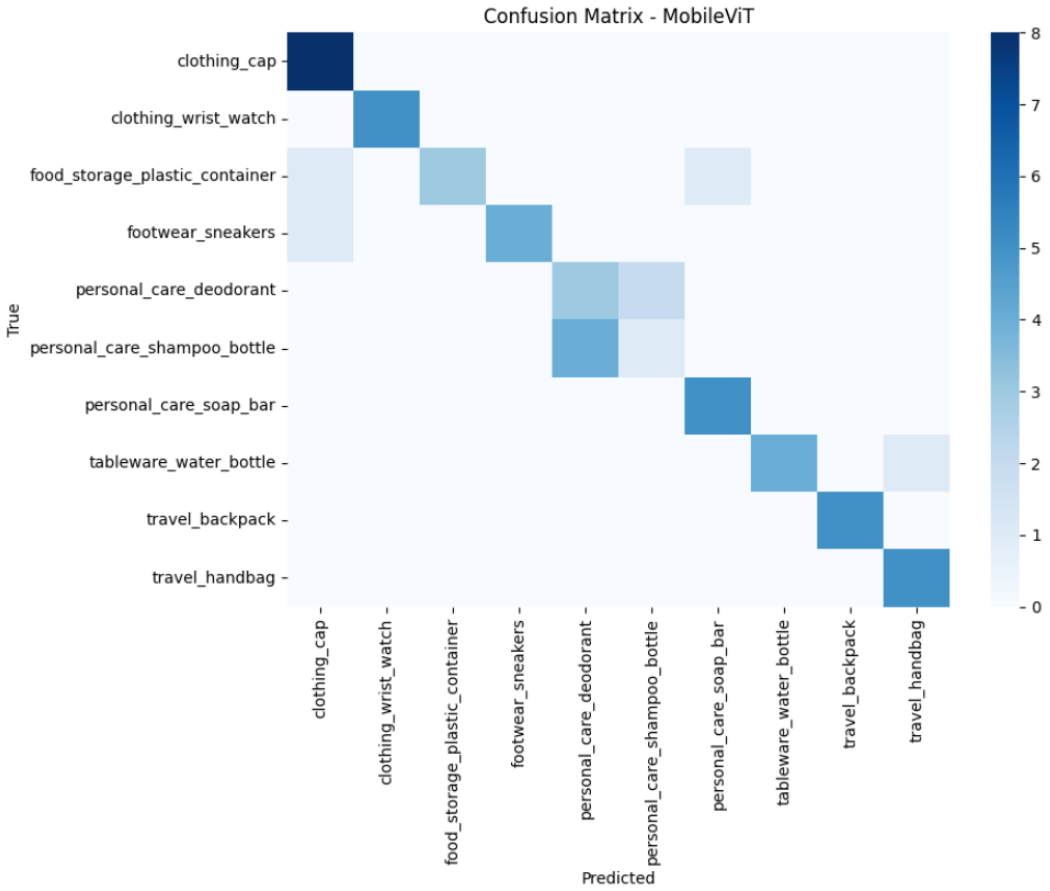
Hence, while Dataset 2 is ~20× larger, its **label entropy and inconsistency** limited learning far more than sample count helped.

Along with these the epochs seem **too low** for dataset 2 and increasing epochs making specific changes to approach and using model with higher parameters will help the training

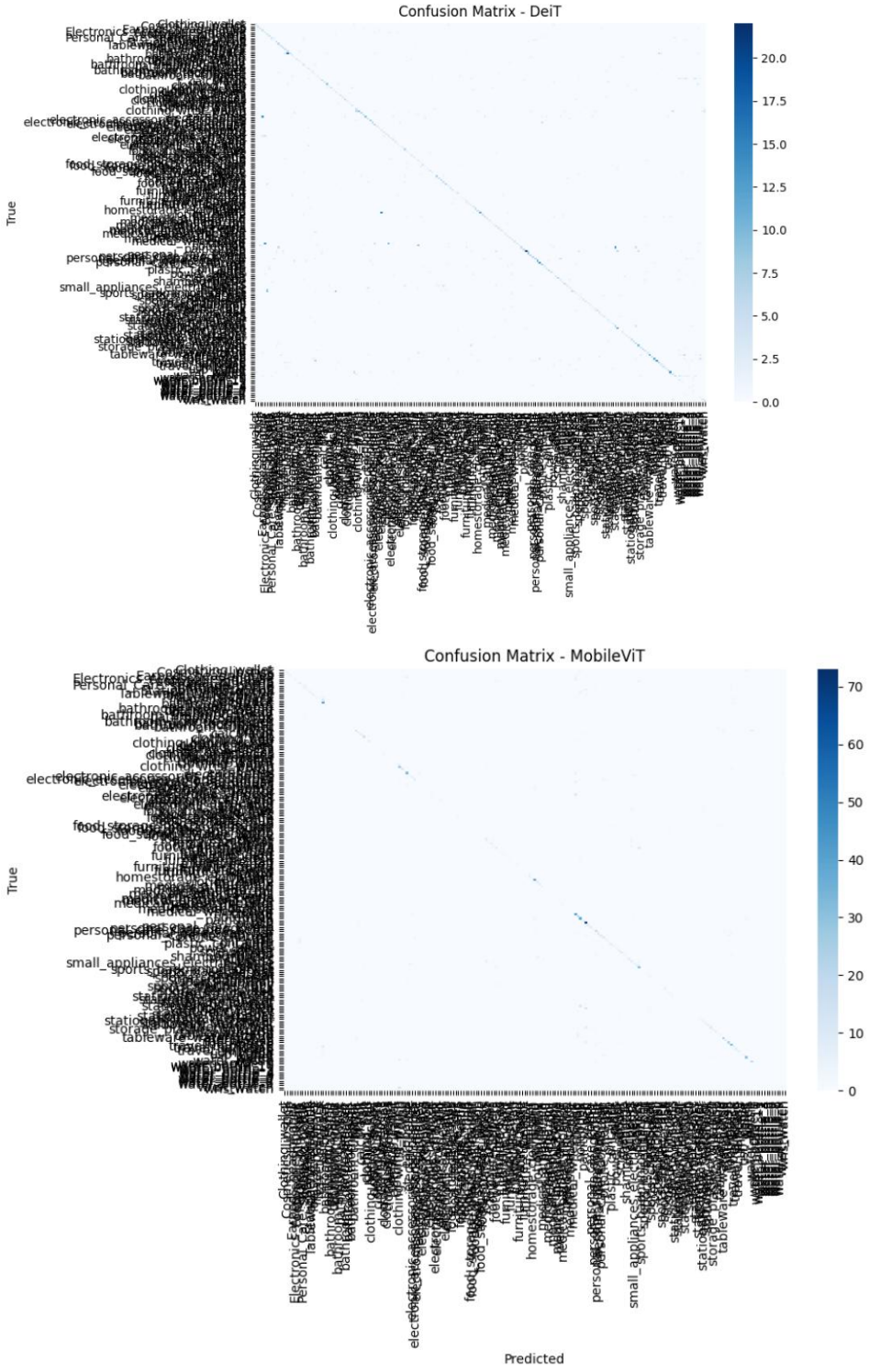
6. Visualization and Error Patterns

- **Confusion Matrix:**


Confusion Matrix - DeiT
True
clothing_cap
clothing_wrist_watch
food_storage_plastic_container
footwear_sneakers
personal_care_deodorant
personal_care_shampoo_bottle
personal_care_soap_bar
tableware_water_bottle
travel_backpack
travel_handbag
7
6
5
4
3
2
1
0



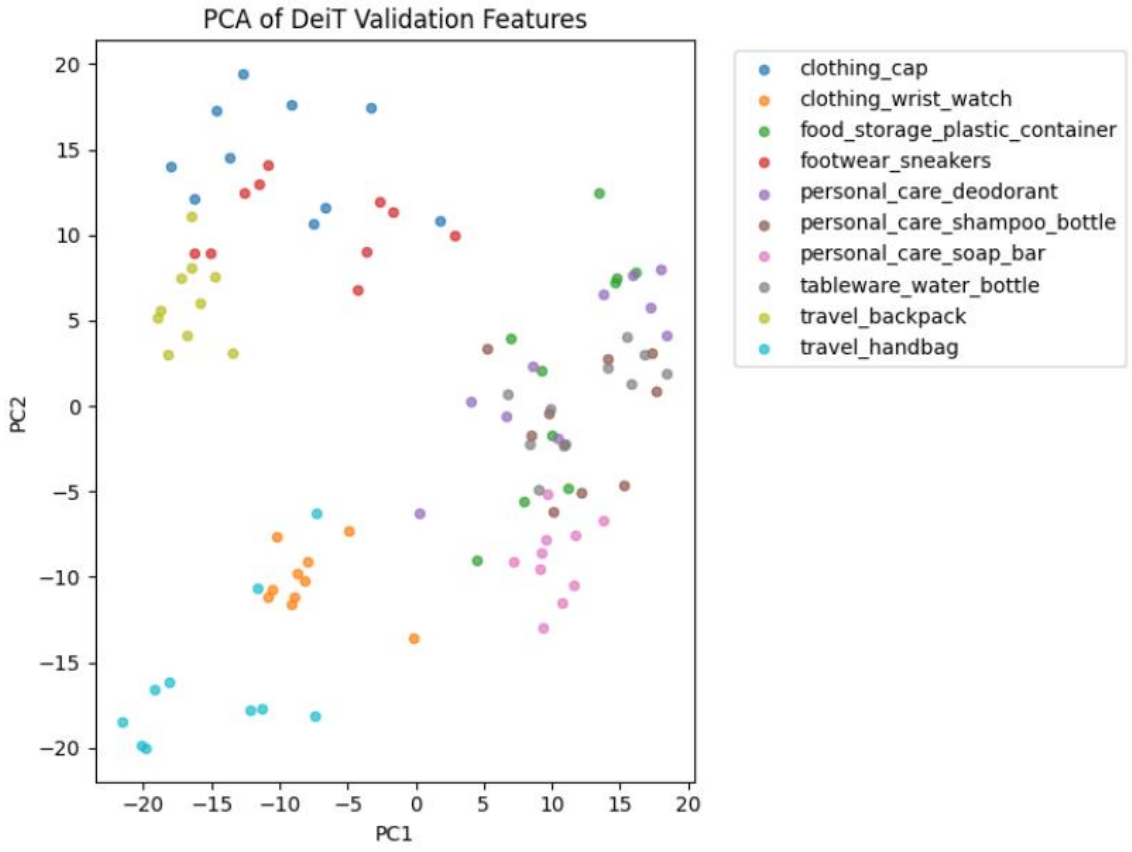
For Dataset 1, most confusion occurred between *shampoo bottle* and *soap bar* (visual similarity).

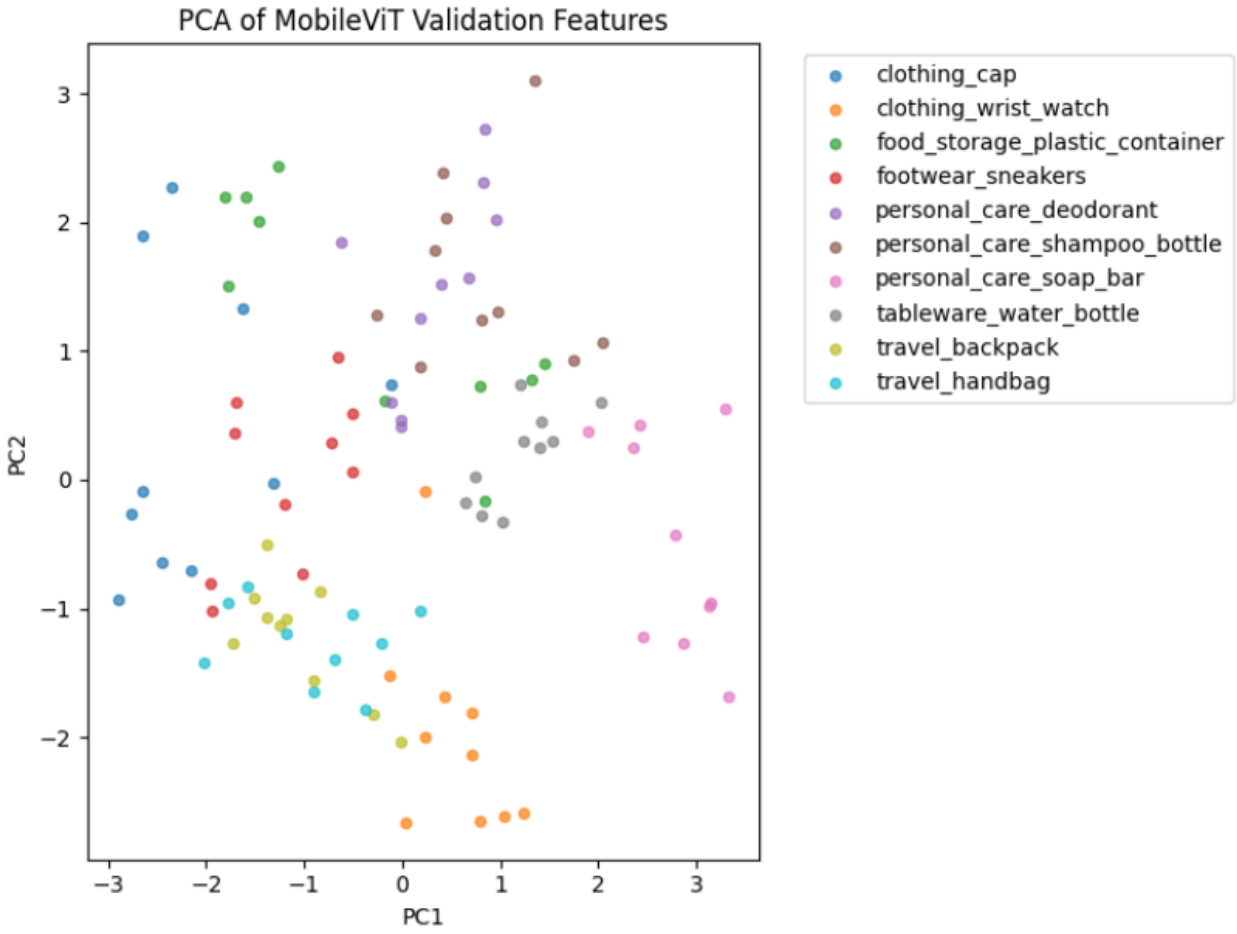


For Dataset 2, high confusion among semantically overlapping classes (*toothpaste* vs *toothbrush*, *wallet* vs *bag*) confirmed noisy label taxonomy.

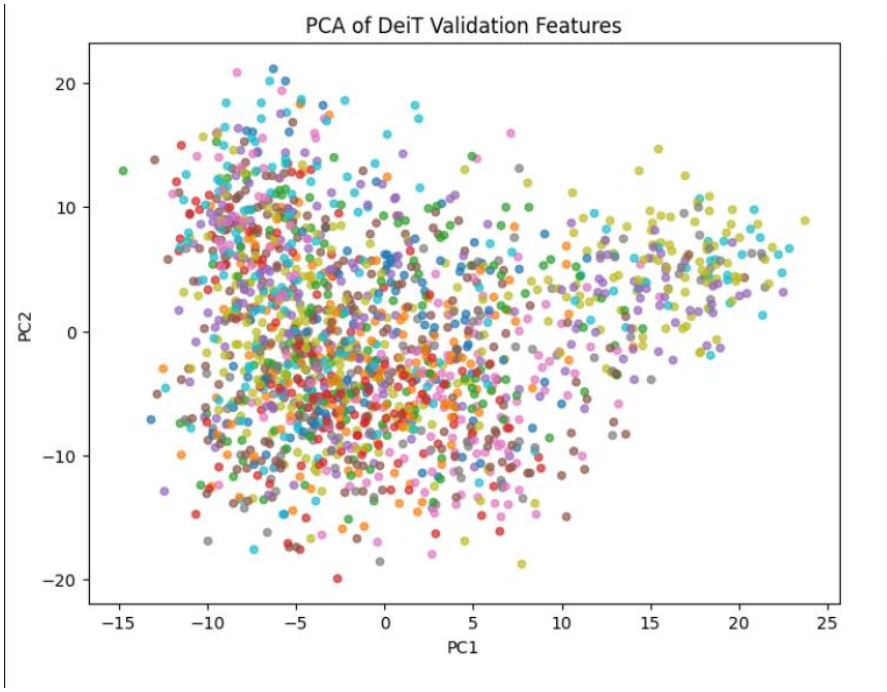
The confusion matrices show that DeiT performs much better than MobileViT. DeiT has a strong, dark diagonal with very few off-diagonal errors, meaning it classifies almost all categories correctly and maintains clear separation between similar classes. In contrast, MobileViT shows more confusion between visually similar items (like shampoo vs. soap or backpack vs. bottle), indicating lower precision. Overall, DeiT demonstrates higher accuracy and generalization, while MobileViT trades some accuracy for efficiency on lighter devices.

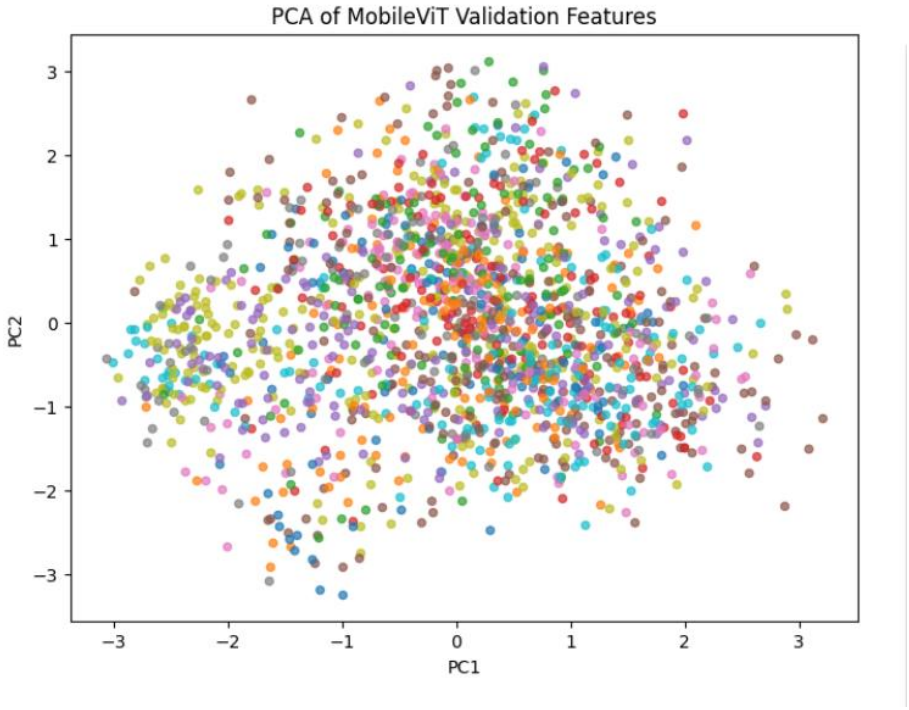
- **PCA Feature Plots:**





Dataset 1 has tight, well-separated clusters although for mobile vit they seem close





Dataset 2 has dense overlaps and diffuse boundaries reflecting semantic noise. This is also a cause of low scores here

- **Retrieval Examples:**
  refer the video for recorded example and this is sample screenshot of the same

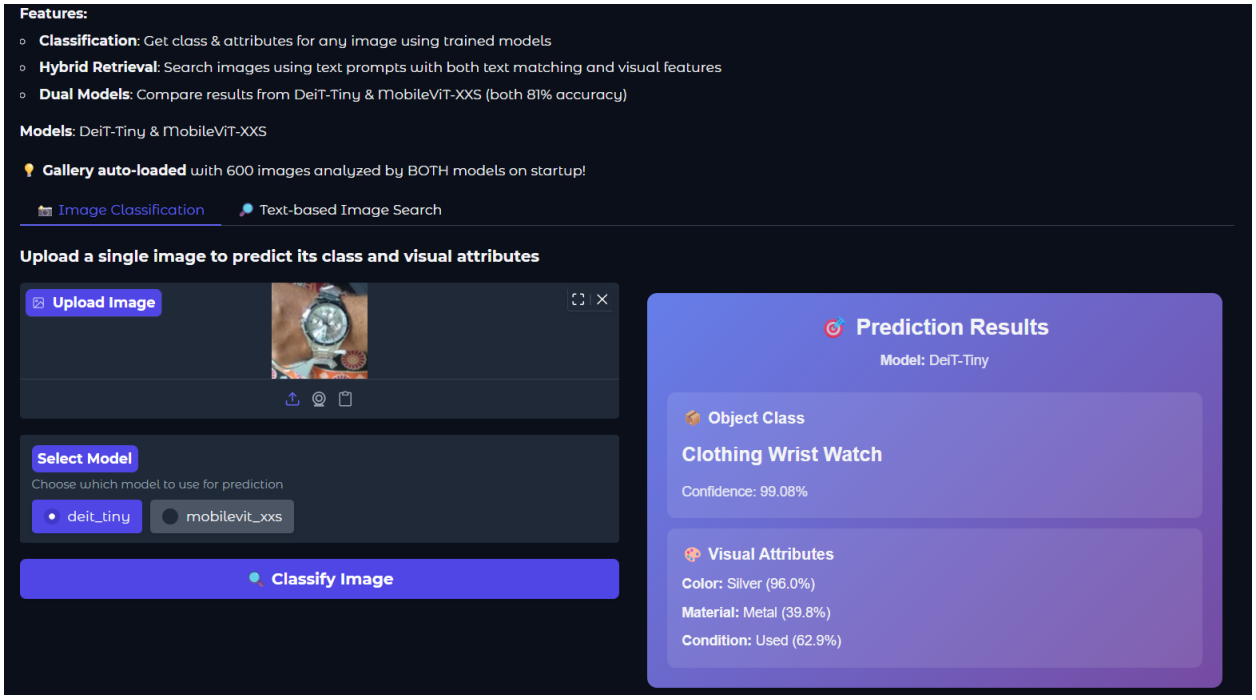
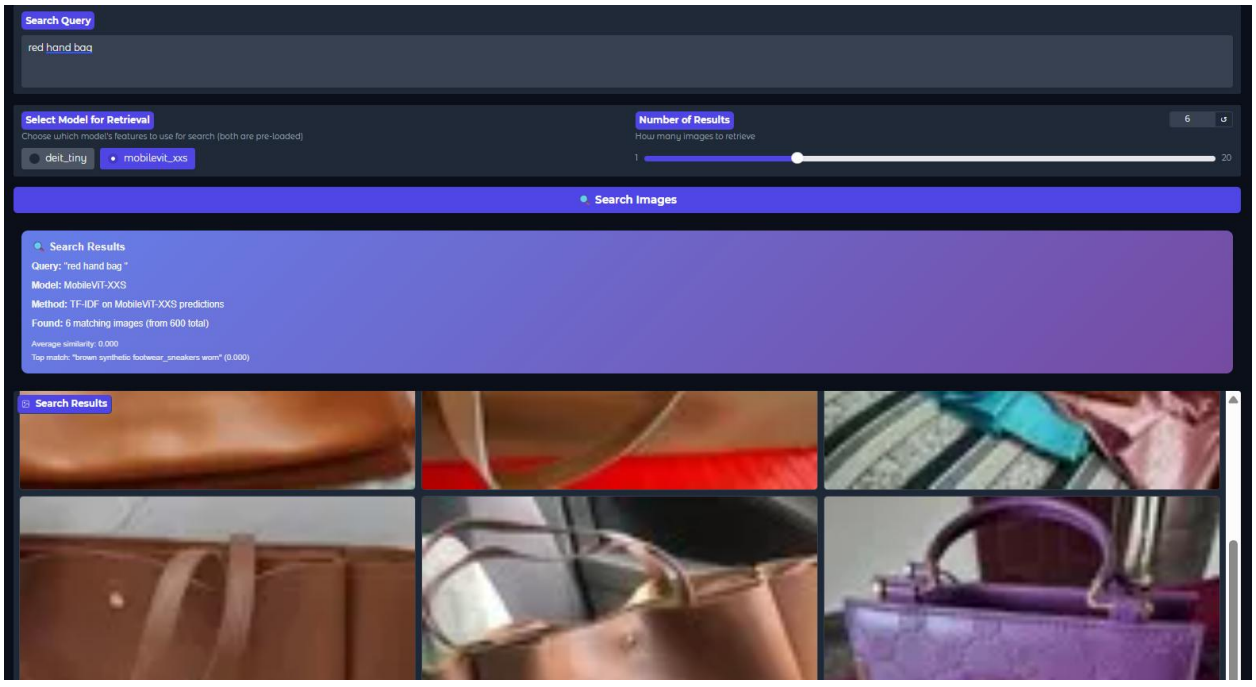
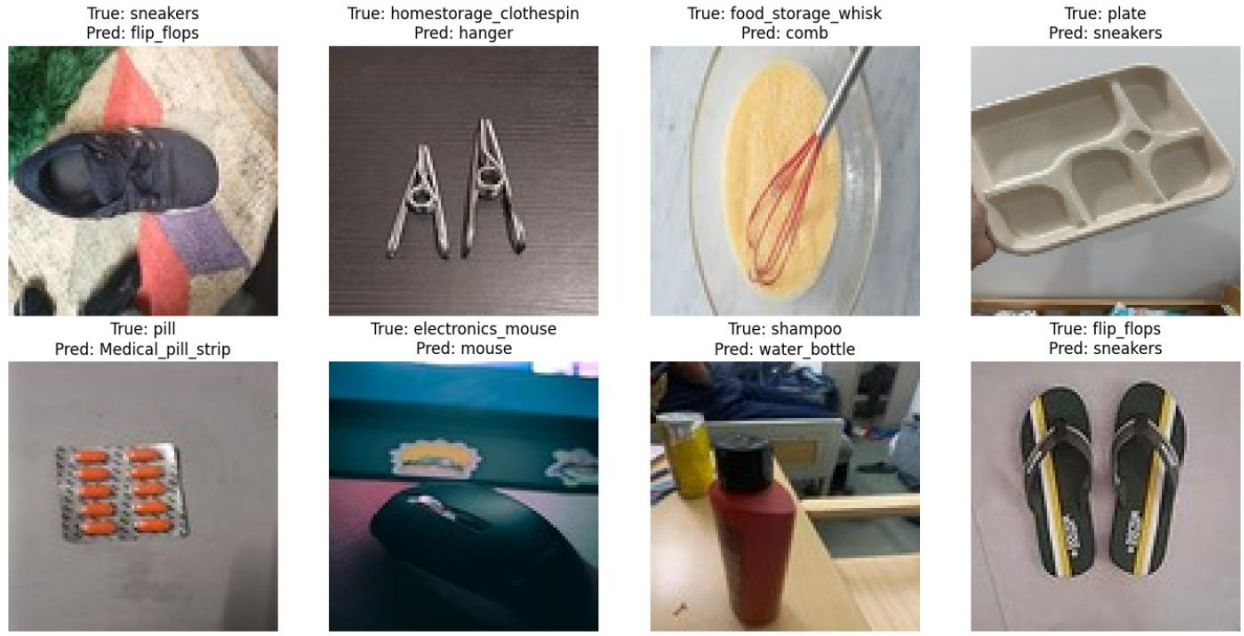


Image classification and attribute prediction example



Text to image retrieval example

Few misclassifications:



7. Key Insights

Larger models like DeiT performed well only when the data labels were clean and consistent. When the labels were noisy, the model tended to overfit. Improving label quality and applying techniques like label smoothing or Mixup can help reduce this issue.

Increasing dataset size did not always improve results. I noticed that poor or inconsistent annotations limited performance, showing that data quality is more important than quantity. Using only verified and consistent samples gave better outcomes.

Fine-tuning the full model made a major difference. Training only the classification head gave less than 30% F1, but full fine-tuning increased it to around 70%. This shows that adapting the pretrained layers to the new dataset is essential.

The dataset had imbalance across classes, which affected the model's ability to recognize minor categories. Using class-weighted loss functions improved the F1 score slightly and gave more balanced predictions.

Retrieval tasks were highly sensitive to attribute accuracy. Clean and consistent attribute labels directly improved the quality of image–text matching. Ensuring proper labeling and structured attributes gave visibly better results.

I also feel the no of epochs is low and with other optimizer and large epochs maybe the dataset2 models can see considerable change in scores