Sequential Clustering Limitation

Current Approach

1. Cluster 1 forms: [A, C, F, K, N]
    - These 5 images are different from each other
    - But no consideration of future clusters
2. Cluster 2 forms: [B, E, H, L, O]
    - These 5 images are different from each other
    - But only chosen from remaining images
    - No consideration of relationship to Cluster 1
3. Last clusters might be less optimal because:
    - Working with leftover images
    - No global optimization across clusters

Example of Potential Issue

Let's say we have images of:

- Cluster 1: [Dog1, Dog2, Dog3, Dog4, Dog5] (all dogs, but different breeds)
- Cluster 2: [Cat1, Cat2, Cat3, Cat4, Cat5] (all cats, but different breeds)

Even though each cluster has internal diversity (different breeds), the clusters themselves are thematically similar (all dogs vs all cats).

Alternative Approaches Could Include:

1. Global Clustering:

```
# Using hierarchical clustering to first split into 20 groups
hierarchical = AgglomerativeClustering(n_clusters=20)
cluster_labels = hierarchical.fit_predict(selected_features)
```

2. Two-Stage Approach:

- First cluster into 20 groups using global information
- Then ensure diversity within each cluster

3. Iterative Refinement:

- Form initial clusters
- Swap images between clusters to maximize both:
    - Within-cluster diversity

- Between-cluster diversity