Decoding movie genres with Vision

A Machine Learning Approach to Predict Movie Genres from Posters

CONTRIBUTORS

Ahmeda Cheick
Richard Lumpi
Averine Sanduku
Francesca Scipioni

DATASCI 281 - COMPUTER VISION

April 16, 2025

U.C. Berkeley School of Information

Project Introduction

OVERVIEW

We assembled a dataset of **29,265 IMDb movie posters** and built a **visual-only pipeline** that combines classic HSV & HOG descriptors with deep embeddings (ResNet50, ViT) for **genre classification**.

CHALLANGE

Posters come in wildly varying resolutions, often carry **multiple genre tags**, and exhibit **severe class imbalance**—making single-label, purely visual prediction **especially difficult.**

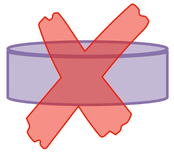
GOAL

Develop a **robust classifier** that distills each poster down to its dominant genre—balancing performance, speed, and implementation complexity across simple models (**Logistic / SVM**) and **fine-tuned CNNs.**

APPLICATION

Enable **smarter content recommendation** and **automated cataloging** by inferring genre information directly from poster art—**no text metadata required**.

Data & Labels Preparation

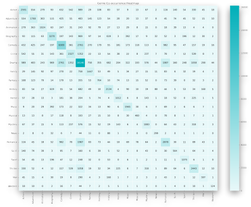


Non-imputable missing values

Non-poster images

53,000+ posters
Year 2000-2024
IMDB rating >7.0

29,265 posters

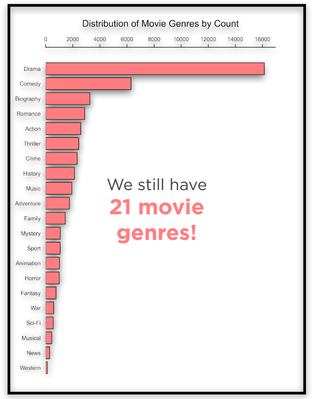


Most movies have **2+ labels**

Movie plots

Movie genre

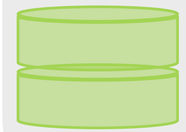
SEMANTIC EMBEDDINGS



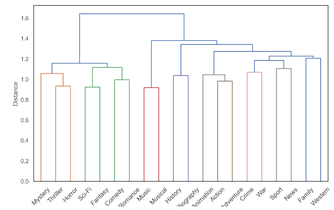
We still have **21 movie genres!**

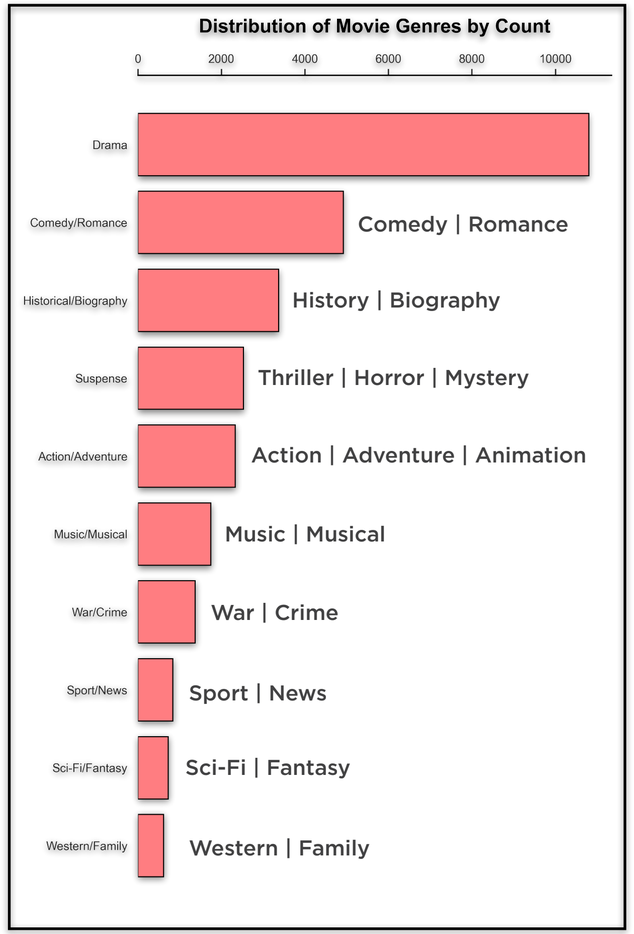
Used **COSINE SIMILARITY** to select **THE GENRE** that mostly matches the movie plot.

29,265 single-labeled posters



Used **HIERARCHICAL CLUSTERING** to reduce labels number





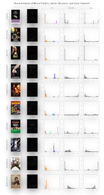
Extracting Features

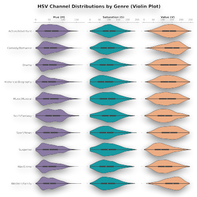
29,265
single-labeled
posters

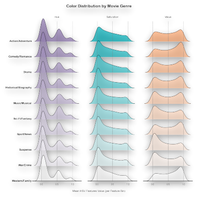
CLASSIC METHODS

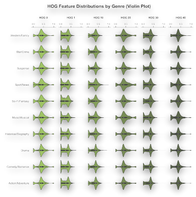
HSV Color Histograms and Moments

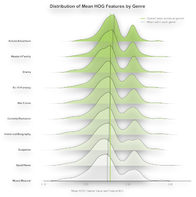
Histogram of Oriented Gradients (HOG)

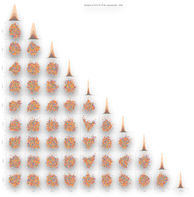


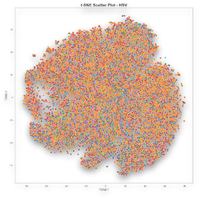


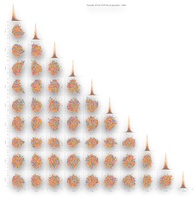


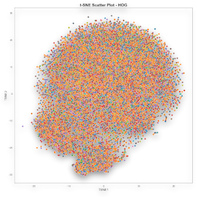








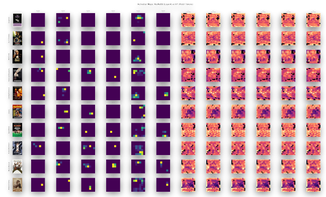


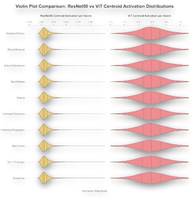


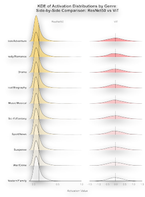
ADVANCED METHODS

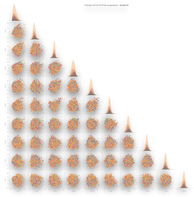
ResNet50 pre-trained CNN

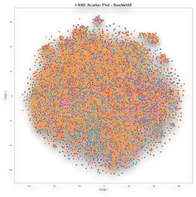
Google's Vision Transformers (ViT)

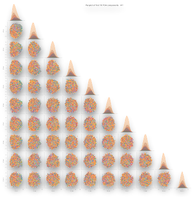


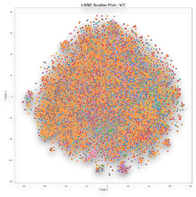








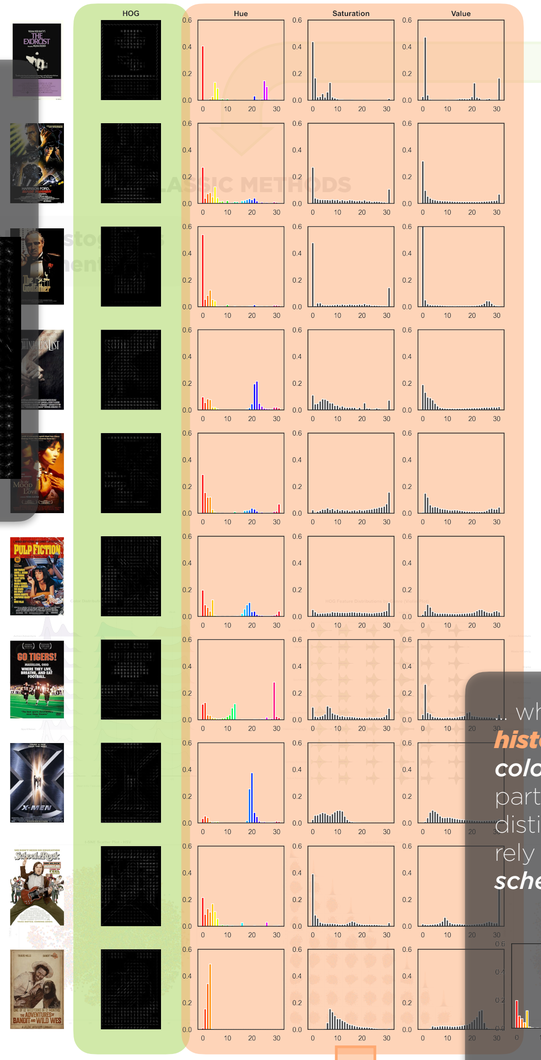




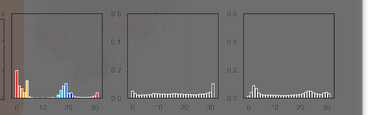
Extracting Features

29,265 single-labeled posters

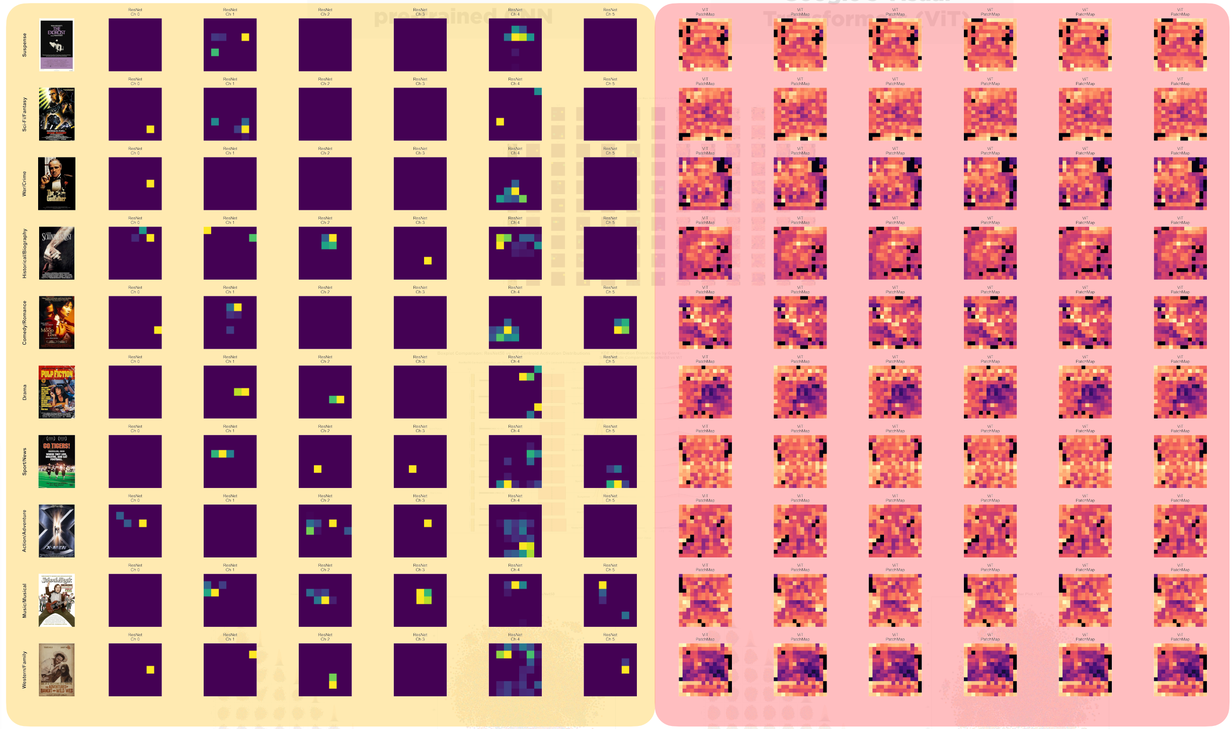
HOG representation strongly responds to ***text*** sections and can ***vectorize patterns*** like human silhouettes...



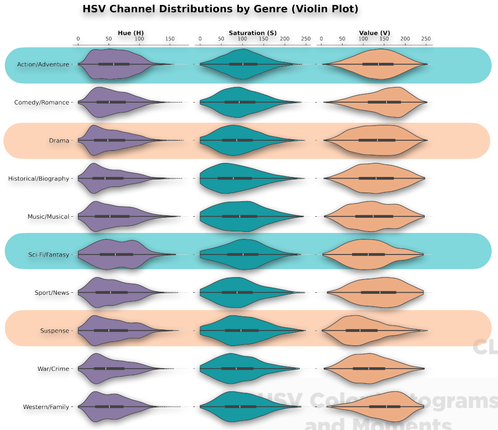
... while ***HSV color histograms*** highligh ***overall color palette,*** making it particularly useful for distinguishing genres that rely on ***distinctive color schemes***.



The ***ResNet50's convolutional layers*** focus on ***local textures and shapes***—bright spots in each patch reflect strong feature responses to edges, text, or distinctive silhouettes

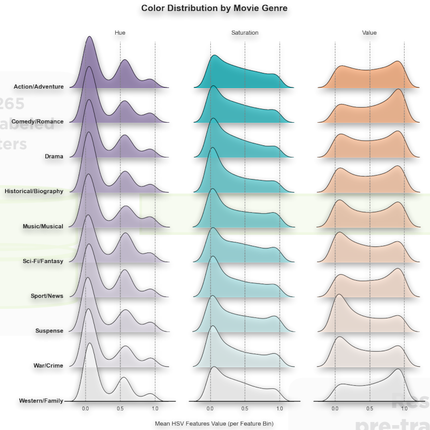


While the ***ViT*** patch tokens capture ***broader contextual relationships*** earlier in the network, using ***self-attention*** to integrate information across the entire poster



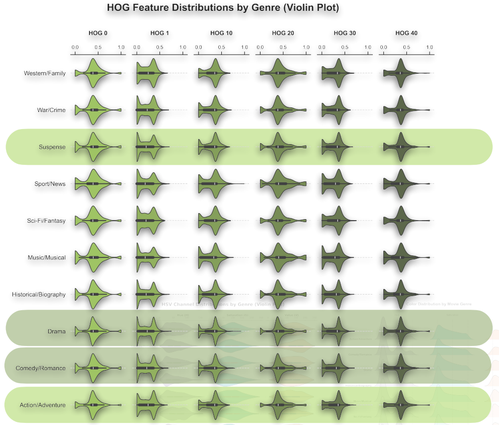
HSV captures ***color distributions:***

genres with ***bright, saturated palettes*** (e.g., Sci-Fi/Fantasy) show distinct ***peaks***, while more ***muted categories*** cluster around ***narrower hue or value ranges.***



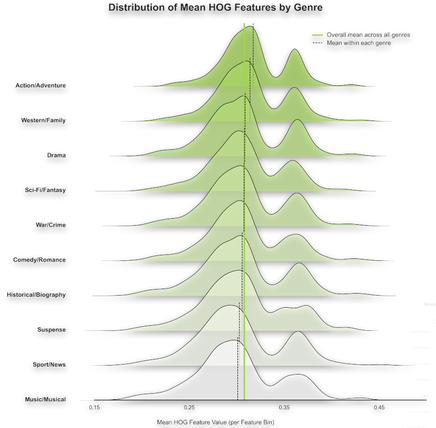
HSV reveals ***clear difference*** in color distributions:

genres such as ***Comedy and Musical*** burst with ***vibrant, saturated hues***, while ***Suspense and War/Crime*** favor ***muted, darker palettes***



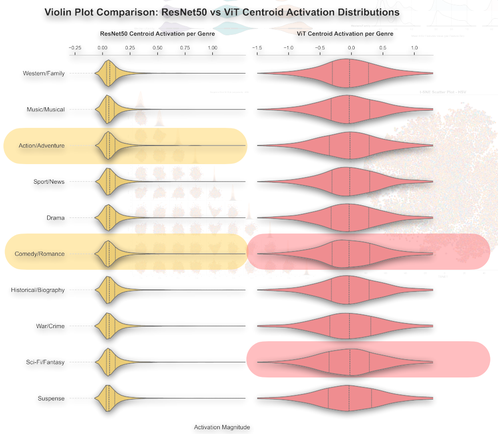
HOG emphasizes **edge and texture** frequencies:

genres with ***busy layouts or strong silhouettes*** (like Action/Adventure) produce ***wider distributions*** in certain HOG bins, whereas *simpler designs* yield ***tighter, more peaked shapes.***



HOG:

posters for ***Action and Adventure*** exhibit a ***wide range of edge intensities***—suggesting busy, detailed layouts—whereas *Romance and Drama* are characterized by ***simpler, more streamlined structural designs***

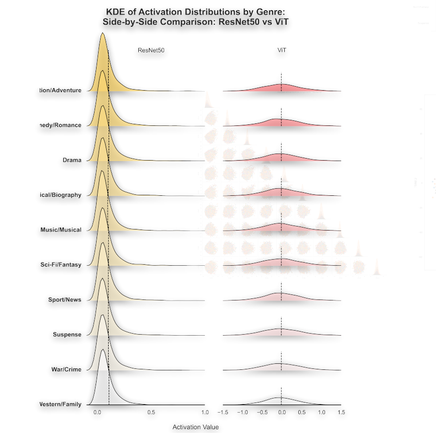


ResNet50:

Captures fine, local textures—busy genres like Action show high activation variance compared to the simpler patterns of Romance.

ViT activations:

Produces uniformly global representations—Sci-Fi posters form ***cohesive structures***, while Comedy posters show more ***scattered activation patterns***.



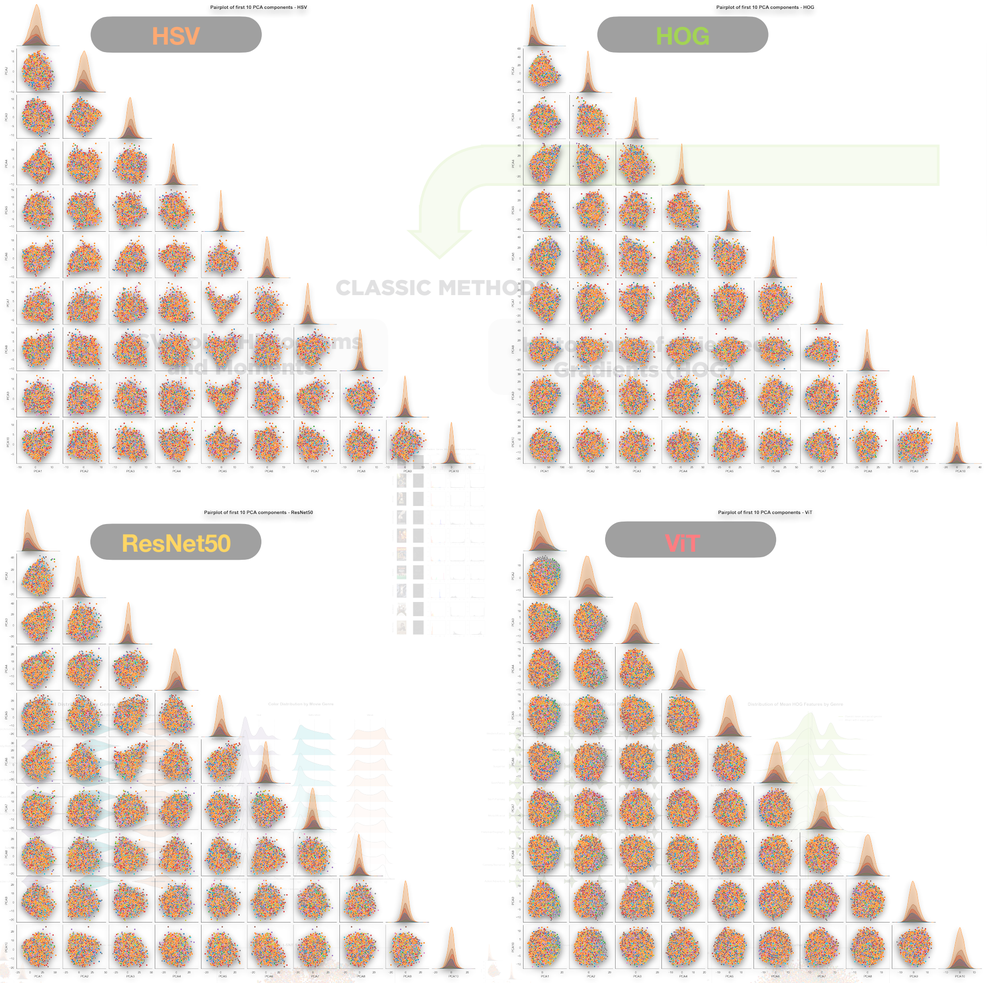
ResNet50:

focuses on local texture details, producing ***sharp activation peaks in visually dense genres***, while ***calmer genres yield lower, more concentrated responses***

ViT activations:

patch-based, global attention strategy generates ***smoother, more uniform activation patterns***, yet still distinguishes genres.

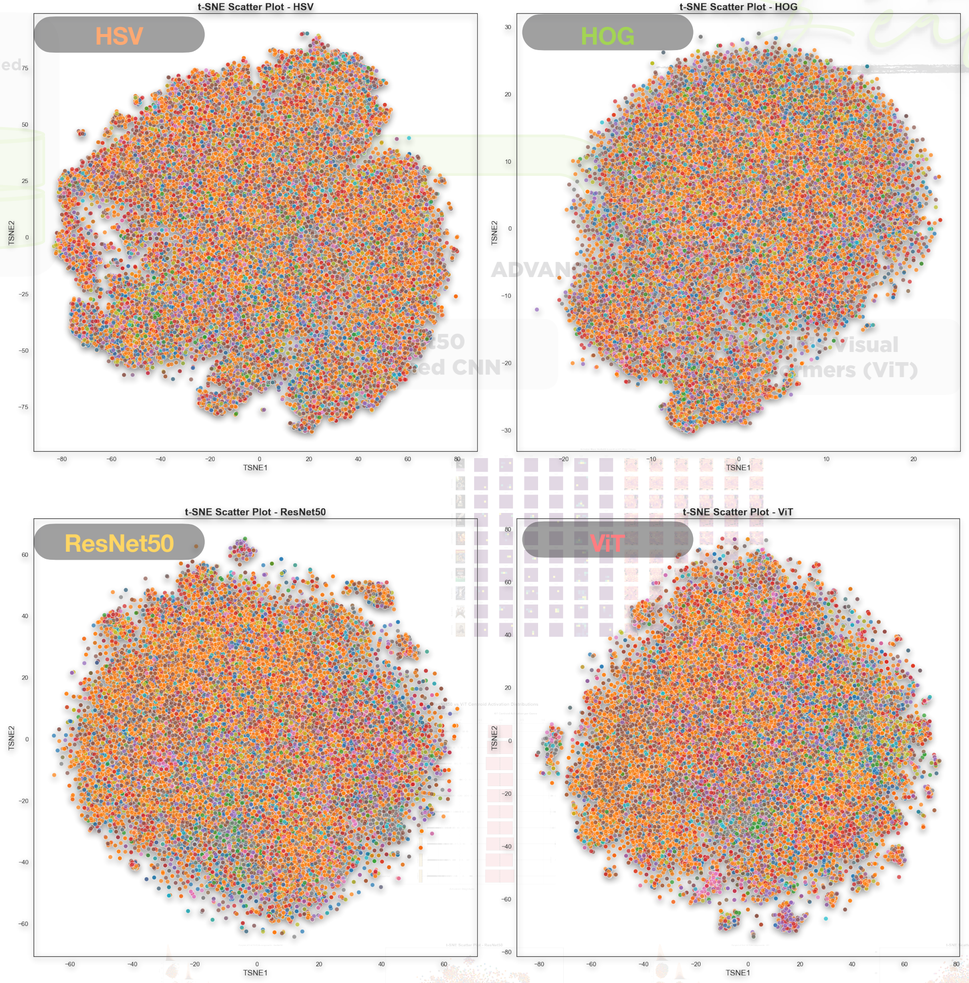
Pair plots of first 10 PCA components



HSV and **HOG** tend to produce ***overlapping clusters***, indicating that their lower-level representations capture ***some common visual characteristics across genres***.

In contrast, the deep features from **ResNet50** and **ViT** exhibit more ***distinct clustering***, suggesting that their learned, high-level representations offer ***better separability*** among movie poster genres.

Scatter plot of t-SNE 2 components



HSV and **HOG** yield relatively ***diffuse clusters*** with considerable overlap, suggesting that these low-level features ***capture shared visual cues*** across genres.

In contrast, the deep features from **ResNet50** and **ViT** form more ***distinct, coherent clusters***, indicating their superior ability to ***capture high-level semantic information*** that differentiates movie poster genres.

Modeling

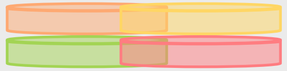
Baseline Model

Predict all classes equally

Baseline Accuracy: 10.%

29,265 posters

Extracted features & their concatenation



MODEL 1
Supporting Vector Machines

Modeling Pipeline

1. **Downsample** DataFrame to balance classes
2. Data Splitting (80/20 Stratified)
3. StandardScaler
4. PCA (85% Variance)
5. **SVC (Hyperparameter Tuning via GridSearch)**

Cross-Validation (StratifiedKFold)

Best Result:

- Best Combination: ViT+HSV
- Validation Accuracy: 30.4%
- Test Accuracy: 25.7%

MODEL 2
Logistic Regression

Modeling Pipeline

1. **Downsample** DataFrame to balance classes
2. Data Splitting (80/20 Stratified)
3. StandardScaler
4. PCA (85% Variance)
5. **Logistic Regression Classifier (Multinomial, tuned on C parameter)**

Best Result:

- Best Combination: HSV
- Validation Accuracy: 17.4%
- Test Accuracy: 15.3%

MODEL 3
for Zero-shot

1. **Downsample** DataFrame to balance classes
2. Load CLIP Model & Preprocessing (ViT-B/32 Variant)

1. Extract Unique Genres & Create **Text Prompts** (e.g., "a movie poster for a <genre> film")
2. **Tokenize** Text Prompts
3. Iterate Over Downsampled Samples
4. Load & Preprocess Image

1. Compute Image Features via **CLIP**
2. Compute **Cosine Similarity** with Text Features & Apply **Softmax**
3. Determine **Predicted Genre** (argmax)
4. Append True & Predicted Labels for **Evaluation**

Best Overall Result:

- Test Accuracy: 39.0%

MODEL 4
ResNet50

1. Define **TARGET IMAGE SIZE (224x224)**
2. Resize-with-Padding Function
3. Load & Preprocess Each Image

1. Load **CSV Data** & **Build Labels**
2. Construct Image (X) & Label (y) Arrays from the CSV
3. Split Dataset with Stratification
4. **Downsample** DataFrame to balance classes

Build Model with ResNet50 Base
Preprocess → ResNet50 → GlobalAveragePooling → Dense → Dropout

1. Wrap Model via KerasClassifier
2. Hyperparameter Tuning using **GridSearchCV (Stratified 5-Fold CV)**

Best Result:

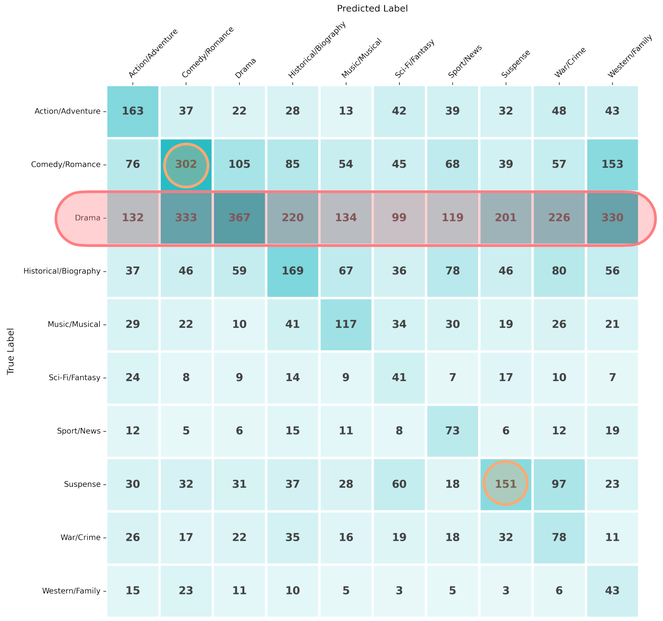
- Accuracy Accuracy: 24.5%

Modeling Results

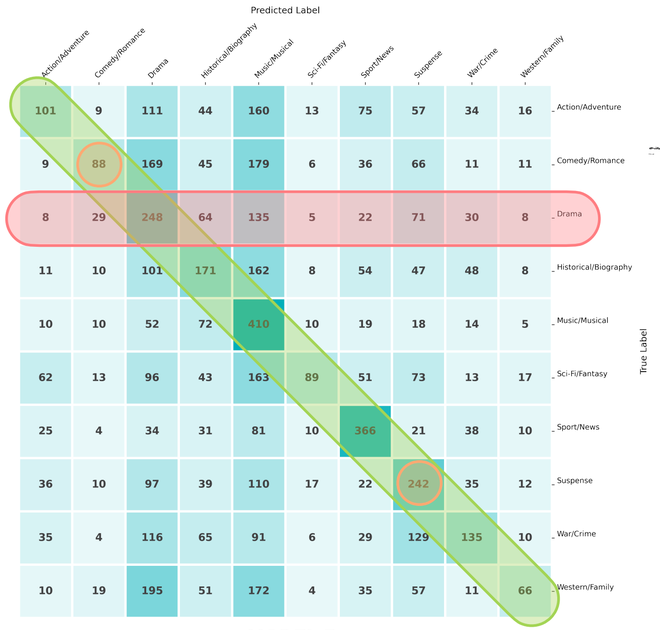
Drama is the most confused class, and the best predicted path the same time

CLIP for Zero-shot is the best performing model

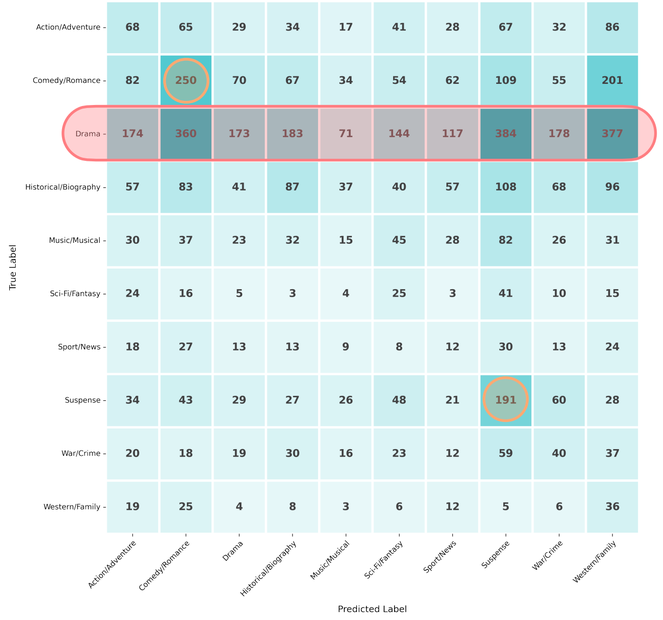
Suspense and **Comedy/Romance** have good predictions across all the models



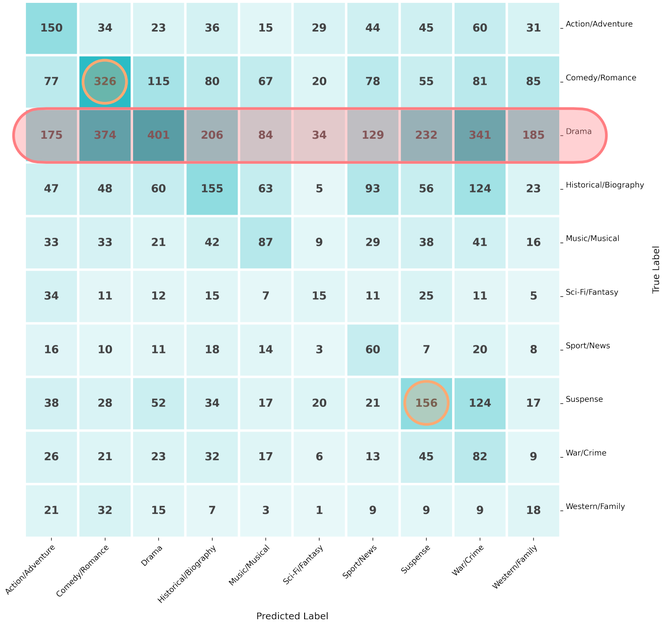
MODEL 1
Supporting Vector Machines



MODEL 3
CLIP for Zero-shot



MODEL 2
Logistic Regression



MODEL 4
ResNet50

Conclusions

KEY QUANTITATIVE INSIGHTS

- **HSV / HOG baselines** extract in milliseconds but plateau at ~15% test accuracy.
- **Best "classical" pipeline** (ViT embeddings + HSV → linear SVM) reaches 26% test accuracy.
- **CLIP for Zero-shot** tops out at **~39% test accuracy** but requires ~20× more compute than the ViT+HSV SVM for equivalent performance.

WHY ACCURACY STALLS

- Purely single-label classification **discards multi-genre cues**—many posters straddle several categories—so recall on minority genres suffers and overall accuracy stagnates.

COMPUTER-VISION NEXT STEPS

1. Multi-Label Training
Align the objective with real poster semantics by predicting all applicable genres.

2. End-to-End Transformer Fine-Tuning
Unfreeze and train larger ViT variants (e.g. ViT-L/16) on poster data for richer, domain-specific representations.

3. Genre-Aware Augmentation
Use adaptive color jitter, random erasing, or GAN-based style transfers to simulate varied poster styles.

4. Domain Adaptation & Metric Learning
Emphasize genre-specific visual cues via adversarial domain alignment or triplet-loss training.

5. Multimodal Fusion
Combine poster imagery with plot embeddings, keywords, or trailer frames to push accuracy beyond 30–40%.

BOTTOM LINE

- Deep embeddings unlock gains, but breaking the 30% barrier on single-genre poster classification will demand richer labels, smarter augmentations, and seamless integration of multiple modalities.