DECODING MOVIE GENRES WITH VISION
A Machine Learning Approach to Predicting Movie Genres from Posters

Ahmeda Cheick, Richard Lumpi, Averin Sanduku, Francesca Scipioni

UC Berkeley School of Information

Abstract

Movie posters are not only marketing tools—they are carefully designed visual compositions intended to communicate key aspects of a film's tone, style, and genre. As such, they serve as a potentially rich source of information for genre classification. In this project, we investigate whether it is possible to predict a film's genre using only its promotional poster image. We experiment with both classical and deep learning approaches, extracting low-level features (HSV histograms, HOG descriptors) and high-level features (from pretrained CNNs and Vision Transformers). We benchmark the performance of various classifiers, including logistic regression, SVMs, a fine-tuned ResNet50, and a zero-shot CLIP model. While handcrafted features achieve modest performance, deep visual embeddings substantially improve classification accuracy. Notably, CLIP achieves the best performance without any training, highlighting the promise of vision-language models for genre prediction. Our findings suggest that while visual cues alone offer a useful signal, incorporating multi-label annotations and multimodal data is key for advancing genre classification systems.

1 Introduction

Movie genres help audiences set expectations and assist streaming services in recommending content tailored to user preferences. While genre labels are typically assigned based on a film's narrative or metadata, humans can often infer a movie's genre simply by glancing at its promotional poster. This intuitive ability is supported by decades of graphic design practice that encodes genre-specific conventions into color palettes, typography, composition, and imagery.

Prior research has demonstrated that low-level visual features—such as gradients, texture, and color distributions—can be used to infer genre information from posters[1]. However, these methods often struggle to generalize beyond simple, cleanly labeled datasets. With the rise of deep learning and large pretrained vision-language models, it is now possible to capture more complex visual semantics and evaluate genre prediction in a more realistic setting.

This project investigates the extent to which movie genres can be predicted from poster images alone, using a dataset of over 29,000 real-world posters with consolidated genre labels. We explore both low-level and high-level visual representations, extracted through classical descriptors (HSV histograms and HOG) and modern deep learning architectures (ResNet50, Vision Transformers, and CLIP). We evaluate performance using traditional classifiers such as logistic regression and support vector machines, as well as a fine-tuned ResNet50 CNN and a zero-shot CLIP model.

By comparing these approaches, we aim to understand the trade-offs between interpretability, computational cost, and predictive power, while shedding light on the feasibility of genre classification as a vision-only task. Our results also emphasize the limitations of single-label genre classification and motivate the need for multi-label and multimodal extensions in future work.

2 Dataset

2.1 Sources

Our dataset combines movie metadata and poster images sourced from two publicly available repositories: the IMDb Non-Commercial Datasets[2] and the OMDb API[3]. IMDb provides a comprehensive listing of movie identifiers, release years, and user ratings, while OMDb complements this with rich

[1] Ivasic-Kos, M., Pobar, M., & Mikec, L. (2014, May). Movie posters classification into genres based on low-level features. In 2014 37th international convention on information and communication technology, electronics and microelectronics (MIPRO) (pp. 1198–1203). IEEE.

[2] https://developer.imdb.com/non-commercial-datasets/

[3] https://www.omdbapi.com

metadata such as movie genres, director and actor names, plot descriptions, and poster URLs.

2.2 Data Collection Pipeline

To build a representative dataset spanning multiple genres and years, we adopted a two-phase data collection approach:

- **Phase 1: 2014–2024 Data** — We began by using a pre-curated dataset containing movies released between 2014 and 2024.

- **Phase 2: 2000–2013 Data** — To improve genre coverage and balance across time periods, we expanded our dataset to include movies released between 2000 and 2013 with IMDb ratings of 7.0 or higher. This yielded a curated list of movie identifiers, which we used to fetch metadata and poster links from the OMDb API.

2.3 Reliability and Representativeness

The OMDb and IMDb datasets are widely used and trusted in academic and industrial settings. While OMDb aggregates its content from IMDb, it enriches the metadata with user-friendly access and additional attributes. IMDb itself is maintained by Amazon and represents one of the most complete sources of film data globally.

To reduce sampling bias, we filtered movies across two decades (2000–2024), ensured a range of genres were included, and prioritized titles with higher user ratings.

3 Exploratory Data Analysis

To understand the structure and quality of our dataset, we conducted an extensive exploratory data analysis (EDA) focused on cleaning, restructuring, and preparing the data for machine learning. The original dataset consisted of over 53,000 movies, each described by metadata fields such as title, year, genre(s), plot summary, IMDb ratings, and URLs to poster images.

3.1 Handling Missing Values

Several metadata fields had missing values, with notable gaps in the `plot`, `poster`, `actors`, and `awards` columns. We applied the following strategies:

- Genres were imputed using IMDb's `title.basics.tsv` dataset.

- Director, actor, and country fields were filled with the placeholder "Unknown".

- Rows missing essential information like plot summaries or posters were dropped.

- IMDb rating and vote counts were filled in using data from `title.ratings.tsv`.

- The `metascore` column was discarded due to excessive missing data.

After these preprocessing steps and the removal of irrelevant entries (e.g., non-movie types), the dataset was reduced to 46,526 clean movie entries.

3.2 Multi-Label Genre Distribution

The movies in the dataset belong to a subset of 26 genres, often being associated with multiple genres. We one-hot encoded the genre list, revealing that 21,122 entries were associated with more than one genre. A genre co-occurrence matrix confirmed common pairings such as *Drama-Comedy* and *Documentary-Biography*, while some genres (e.g., *Western*, *Talk-Show*) were underrepresented. In a first step to reduce the genres for classification we removed *Game-Show*, *Reality-TV*, *Short* and *Talk-Show* to improve class balance. Furthermore, *Documentary* was identified as an ambiguous genre. A documentary poster can easily be mistaken for that of a regular movie. We will delete all movies where *Documentary* is the only genre and remove the word "Documentary" from movies that have it as one of multiple genres. This step resulted in a dataset of 34,493 entries.

3.3 Poster Quality Control

We observed that not all the images linked in the dataset were real movie posters. To ensure visual consistency, we:

- Verified that each poster image was valid and readable.

- Calculated the height-to-width ratio for each image.

- Filtered out any image with an aspect ratio outside the range [1.3, 1.7], which typically indicates non-poster images.

This process further refined the data set to 29,265 valid movie entries with clean posters.

3.4 Single-Label Mapping Using Plot Semantics

Inherently ambiguous, our aim is to investigate if a poster possesses styles and characteristics that define a specific genre. To do so, we aim to collapse the multi-label posters into a single-genre classification problem. For posters with more than one genre associated, we based the label on plot content. We used OpenAI's `text-embedding-3-large` model to generate semantic embeddings of movie plots and compared them to genre embeddings using cosine similarity. For each movie, the genre most semantically similar to its plot was selected as the primary label.

3.5 Genre Grouping via Clustering

To address long-tail distribution issues and reduce model complexity, we grouped semantically similar genres into macro-categories using hierarchical clustering. This process yielded a consolidated set of genre groups, whose distribution is shown in Figure 1:

- **Suspense**: Thriller, Horror, Mystery
- **Sci-Fi/Fantasy**: Sci-Fi, Fantasy
- **Comedy/Romance**: Comedy, Romance
- **Action/Adventure**: Action, Animation, Adventure
- **Historical/Biography**: History, Biography
- **Music/Musical**: Music, Musical
- **War/Crime**: War, Crime
- **Western/Family**: Western, Family
- **Sport/News**: Sport, News
- **Drama**: Standalone due to high volume

3.6 Final Dataset

The final dataset consists of **29,265** movie entries, each with a validated poster image, cleaned metadata, and a single consolidated genre label suitable for classification. This refined dataset forms the foundation for the feature extraction and modeling pipeline presented in the next sections.

4 Feature Extraction

This section presents two complementary strategies for extracting visual features from movie posters to support genre classification. The first approach relies on classic image descriptors, namely HSV color statistics and Histogram of Oriented Gradients (HOG). The second leverages advanced deep learning models, specifically, the pretrained ResNet50 convolutional network and a Vision Transformer (ViT), to capture semantic and structural information. Feature vectors from all methods were saved to disk for downstream analysis and modeling.

4.1 Classic Feature Extraction

Classical feature extraction techniques offer a compact representation of low-level image attributes. In this study, HSV color histograms and Histogram of Oriented Gradients (HOG) were employed to examine variations in hue, saturation, and brightness, along with local texture information present in the posters. These features were selected based on the hypothesis that they encapsulate the immediate visual impression or "glimpse" that a viewer perceives when first encountering a poster.

4.1.1 HSV Color Histograms and Moments

HSV histograms decompose the image into hue, saturation, and value (brightness) distributions, while color moments (mean, standard deviation, and skewness) capture statistical properties like dominant color palette or brightness. In particular, we decomposed each poster into three normalized 32 bin histograms. A visual breakdown of the average channel values per poster across genres is shown in Figure 2, representing the dominant color family (Hue), the intensity of the colors (Saturation) and the brightness (Value), hinting at genre-specific patterns. For example, *Comedy/Romance* and *Western/Family* genres appear to have brighter movie posters, while *Suspense* movies exhibit lower value scores, consistent with darker poster imagery.

To visualize the distribution of HSV features more granularly, we plotted violin plots of each HSV channel across genres (Figure 3). These plots reveal broader variances in dominant colors (Hue) and color saturation within genres like *Action/Adventure*, and more constrained distributions in genres like *Drama*. Further we computed average HSV histograms per genre and visualized their distributions using smoothed kernel density estimates (Figure 4).

While visual inspection indicates some variability among genres across the three channels, it does not reveal a clear separation, aside from certain brightness patterns that still appear to overlap across categories. Overall, distinguishing consistent and distinctive patterns across the ten genres through visual examination of the features proved challenging.

4.1.2 Histogram of Oriented Gradients (HOG)

Histogram of Oriented Gradients (HOG) captures local gradient structures and edge information, thereby encoding aspects of shape and texture. For feature extraction, each poster was resized to a resolution of 400×300 pixels, and a sliding window of 16×16 pixels with 8 orientation bins was applied. This resulted in 450 cells, each represented by an 8-dimensional vector, yielding a final feature vector of length 3,600. Posters from various genres were analyzed, and both the original images and their corresponding HOG representations were visualized alongside the HSV histograms (Figure 5). Visual inspection indicates that the HOG features are particularly responsive to textual regions and are capable of capturing structural patterns, such as human silhouettes.

Violin plots of selected HOG features (Figure 6) further demonstrate that genres exhibit distinct gradient-based texture patterns. For instance, genres such as *Action/Adventure* and *Sci-Fi/Fantasy* display broader distributions with higher density in HOG feature bins 0 and 10, suggesting richer edge content and more varied gradient orientations. In contrast, genres like *Drama* and *Historical/Biography* exhibit more compact and symmetric distributions across the selected HOG features, indicating smoother textures and less directional variation. The clear divergence in distribution shapes and spreads across genres highlights the discriminative power of low-level gradient information for genre classification.

To investigate the degree to which HOG features can differentiate between genres we computed kernel density estimates (KDEs) of the mean HOG magnitudes for each genre and visualized their distributions in Figure 7. These plots reveal potential systematic differences across genres. For example, genres such as *Action/Adventure* and *Western/Family* exhibit relatively higher overall HOG intensities, aligning with the intuition that these genres often feature dynamic and texturally complex visual elements (for example explosions). In contrast, genres like *Music/Musical* and *Sport/News* display lower HOG magnitudes, indicative of smoother, less structurally complex imagery. These patterns suggest that HOG descriptors might capture some genre-specific structural cues.

Figure 8 shows the deviation of average HOG orientation histograms per genre from the global mean, capturing how strongly each genre emphasizes or suppresses specific gradient directions. Each bar represents the difference between a genre's average magnitude at a given orientation and the overall average across all genres. Positive deviations (in green) indicate an overrepresentation of that orientation in a genre's poster textures, while negative deviations (in red) indicate underrepresentation. For instance, genres such as *Action/Adventure* and *Western/Family* show strong positive deviations across a wide range of orientations, suggesting highly varied and directional gradient structures. In contrast, genres like *Music/Musical* and *Suspense* exhibit consistently negative deviations, implying smoother textures or less pronounced edge directions. This orientation-based analysis highlights how genres differ not only in the presence of gradients, but also in their angular preferences, potentially reflecting stylistic conventions in poster design and font choices.

4.2 Advanced Feature Extraction

In addition to low-level, handcrafted features, we used deep visual representations extracted from pretrained ResNet50 and Vision Transformer (ViT) models to capture high level semantic information from movie posters. While low level features such as color histograms and HOG descriptors were chosen to capture "glimpse" impression, they may lack the discriminative power necessary to differentiate between genres. In contrast, deep features derived from ResNet50 and ViT are capable of encoding richer contextual cues, such as compositional arrangements, color interactions, and spatial structures that might be characteristic of specific genres.

For both deep learning models poster images were resized to 224×224 pixels, with padding applied to preserve the original aspect ratio. Feature embeddings were then extracted, yielding a 2,048-dimensional vector from the ResNet50 model and a 768-dimensional vector from the ViT model.

To illustrate the increasing abstraction of features captured by these models, Figure 9 presents intermediate activation maps from both ResNet50 and ViT for ten representative movie posters. The ResNet50 activations were taken from the final convolutional block (layer4), while the ViT visualizations represent spatially arranged patch tokens, averaged across feature dimensions. ResNet50 tends to produce highly localized activations, often highlighting central objects or prominent visual elements such as faces or weapons. In contrast, ViT activation maps exhibit more diffuse and abstract patterns, reflecting the model's global receptive field and attention-based processing. These differences underscore the complementary nature of the two architectures: ResNet50 emphasizes salient local features, whereas ViT is more attuned to holistic, scene-level semantics.

Figure 10 presents side-by-side violin plots of the per-genre centroid activations for ResNet50 (left) and ViT (right). Each violin shows the full distribution of average activations across all posters in a genre, with the dashed lines marking the 25th, 50th, and 75th percentiles. ResNet50 activations are strongly skewed towards higher magnitudes—especially in genres like *Action/Adventure* and *War/Crime*—indicating that its convolutional filters fire intensely on localized edges and text. In contrast, ViT activations form broader, more symmetric shapes centered near zero, reflecting its diffuse, attention-based aggregation of global context.

Figure 11 shows kernel density estimates (KDEs) of the same centroid activations, again comparing ResNet50 (left) against ViT (right) for each genre. The ResNet50 curves feature sharp peaks at low activation values and long right tails, highlighting occasional but strong filter responses. ViT densities, by contrast, exhibit smooth, bell-shaped profiles of moderate width, underscoring the transformer's tendency to distribute attention more evenly across the entire poster. Together, these plots illustrate how ResNet50 emphasizes salient local structures while ViT captures holistic, scene-level semantics.

4.3 PCA and tSNE Visualisation

To further explore whether visual layout features contribute to genre differentiation, we conducted a dimensionality reduction analysis using Principal Component Analysis (PCA). Specifically, feature vectors derived from HSV histograms, HOG descriptors, and the deep representations from ResNet50 and ViT were projected into a 10-dimensional orthonormal basis. This allowed for visual inspection of potential genre-specific clustering in a lower-dimensional space. While PCA is particularly suited to preserving general structural information (for example prominent edges in the case of HOG features), it inherently sacrifices fine-grained detail. As a result subtle distinctions are often lost.

An examination of pairwise projections of the 10 principal components for each feature vector (Figure 12) reveals no clear clustering or separation by genre, suggesting at first sight that no genre-specific visual structures are preserved.

In addition, we applied t-distributed Stochastic Neighbor Embedding (t-SNE) to investigate potential non-linear separability within the feature space. Projecting the same set of high-dimensional feature vectors into two dimensions using t-SNE did not yield any visually discernible genre-based clustering either, indicating that a low dimensional non-linear embedding fails to expose genre-specific structures in the extracted visual features.

5 Model Development

To assess the effectiveness of both low-level and high-level visual representations in predicting movie genres, we explored three primary modeling strategies: (1) traditional classifiers such as Logistic Regression and Support Vector Machines (SVM), (2) a fine-tuned ResNet50 convolutional neural network (CNN), and (3) a zero-shot classification approach using the CLIP model.

5.1 Traditional Classification with Extracted Features

In the traditional pipeline, each movie poster was first transformed into a feature vector using one or more of the four extraction techniques: HSV color histograms, HOG descriptors, ResNet50 embeddings, or ViT embeddings. These feature vectors were standardized and passed through a PCA transformation to reduce dimensionality while preserving 95% of the variance. Two classifiers were tested: Logistic Regression and SVM.

To mitigate class imbalance, we undersampled the training data: each genre class was reduced to match the size of the smallest class, resulting in a balanced training set of 4,960 entries. Stratified 5-

fold cross-validation was used for hyperparameter optimization. Models were evaluated on the test set using overall accuracy, classification reports (precision, recall, F1-score), and confusion matrices.

Logistic regression models exhibited lower variance and reduced susceptibility to overfitting in comparison to SVMs. While SVMs consistently attained higher accuracy on the training data, their generalization performance on the test set remained comparable to that of the logistic regression models. This pattern indicates limited transferability beyond the training environment and suggests that the collapsing of genre categories may have introduced intrinsic ambiguity into the classification task(Table 1).

5.2 Fine-tuned ResNet50 CNN

To investigate whether an end-to-end deep learning approach could outperform traditional models, we fine-tuned a ResNet50 model pretrained on ImageNet. The original classification head was replaced with a Global Average Pooling layer, followed by a dense layer with 256 units, a 30% dropout layer, and a final softmax output layer with 10 units corresponding to the genre classes. While the base model was partially frozen to retain generalizable low-level features, the newly added layers were trained using the entire dataset, with class weighting applied during training to compensate for class imbalance.

The model was trained for up to 10 epochs using early stopping based on validation loss. Despite the architectural depth and capacity of ResNet50, the CNN did not surpass the best-performing classical model. It achieved a test accuracy of 25%, which is comparable to the performance of the Logistic Regression and SVM classifiers. These findings support the hypothesis that genre classification from poster images is limited by intrinsic visual and labeling ambiguities, reducing the dataset's discriminative potential.

5.3 Zero-Shot Classification using CLIP

In addition to supervised models, we explored OpenAI's CLIP model (ViT-B/32 variant) for zero-shot classification. CLIP was used to compare image embeddings with text embeddings generated from genre prompts (e.g., "a movie poster for a drama film"). The genre associated with the highest similarity score was selected as the predicted label. This approach required no training or labeled data, offering a scalable and efficient alternative.

When evaluated on a stratified test set, the CLIP model achieved a test accuracy of 31.71%. This outcome highlights the effectiveness of pretrained vision-language models, even without domain-specific fine-tuning. Although it slightly outperforms previous models, the test accuracy remains broadly consistent with earlier results, further reinforcing concerns about the inherent challenges and limitations associated with the dataset.

5.4 Evaluation and Confusion Matrix Comparison

Table 1 summarizes the four models based on their weighted average precision, recall, and F1 score evaluated on the test set. The CLIP Model performs best overall, achieving the highest precision (0.42), recall (0.32), and F1 score (0.32). This suggests it's the most balanced and effective model across the board.

Figure 13 compares confusion matrices across the four modeling approaches. While all models exhibit difficulty in achieving strong diagonal dominance in their respective confusion matrices, the CLIP model shows a notable improvement in generalization which is mostly visible in the dominant class Drama. Logistic Regression, SVM and the ResNet50 model all suffer from significantly low recall on this dominant class. In contrast, the CLIP model achieves a considerable improvement in recall. Comprehensive classification reports are provided in the accompanying notebook.

5.5 Efficiency, Accuracy, and Future Directions

Our results highlight an important trade-off. Classical models are fast and interpretable but limited in expressiveness. Deep learning models, like fine-tuned CNNs, offer richer representations but at increased computational cost. Meanwhile, zero-shot CLIP delivers strong out-of-the-box performance without any training, pointing to the promise of multimodal pretraining.

Despite employing various modeling techniques, including some of the latest machine learning architectures, the decline in performance on unseen data, coupled with persistent imbalances between precision and recall (particularly in the dominant class Drama) reinforces concerns related to class imbalance, labeling ambiguity, and limited visual separability between classes. Even when trained on balanced or class-weighted datasets, the models consistently demonstrate a limited ability to disentangle genre-specific visual cues from poster images, leading to poor generalization and further emphasizing the inherent ambiguity within the dataset. These findings indicate that future research should prioritize:

- Multi-label classification strategies to better capture genre co-occurrence.
- Augmenting visual features with plot summaries and trailers in a multimodal framework.
- Investigating larger or fully fine-tuned ViT models and CLIP variants.
- Applying self-supervised learning or contrastive losses specific to genre semantics.

In summary, while no approach fully overcomes the genre classification barrier, pretrained vision-language models like CLIP offer a promising path forward.

A Tables

Table 1: Selected Results for Different Feature Combinations of Train, Validation, and Test Accuracies (LogReg / SVM).

Feature Set	**Train Acc**	**Val Acc**	**Test Acc**
HSV	0.21 / 0.42	0.17 / 0.17	0.15 / 0.16
HOG	0.40 / 0.83	0.14 / 0.15	0.13 / 0.13
ResNet50	0.38 / 0.78	0.26 / 0.26	0.23 / 0.23
ViT	0.43 / 0.90	0.30 / 0.30	0.25 / 0.26
HOG+HSV	0.71 / 1.00	0.16 / 0.16	0.14 / 0.15
ResNet50+HOG	0.53 / 0.90	0.25 / 0.25	0.21 / 0.21
ViT+HSV	0.45 / 0.81	0.30 / 0.30	0.25 / 0.26
ResNet50+ViT	0.43 / 0.82	0.30 / 0.29	0.25 / 0.26

Table 2: Summary Table comparing the four models based on their weighted average precision, recall, and F1 score.

Model	**Precision**	**Recall**	**F1**	**Support**
LogReg	0.38	0.25	0.27	5853
SVM	0.38	0.26	0.27	5853
ResNet50	0.35	0.26	0.27	5853
CLIP	0.42	0.32	0.32	5853

B Figures

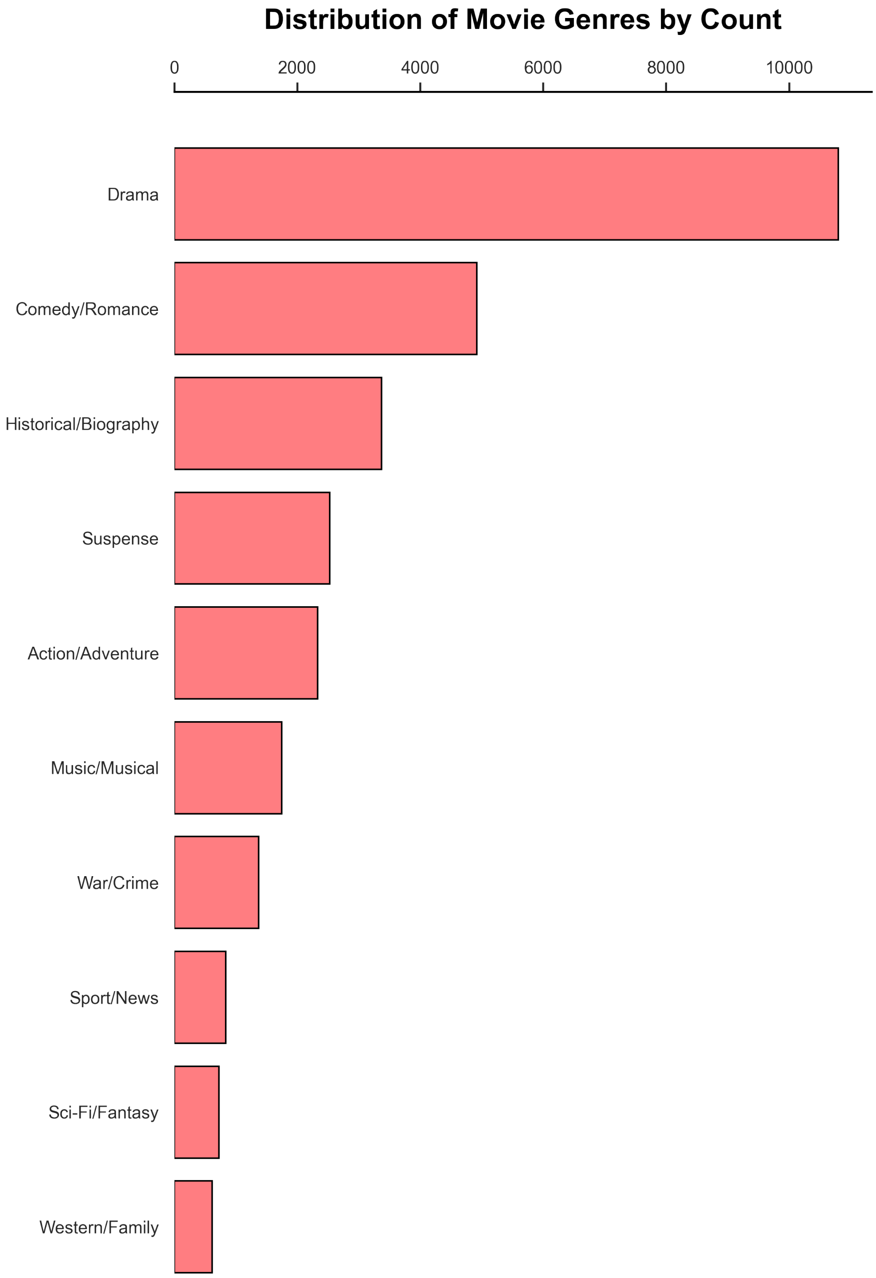

Figure 1: Distribution of Movie Genres by Count.

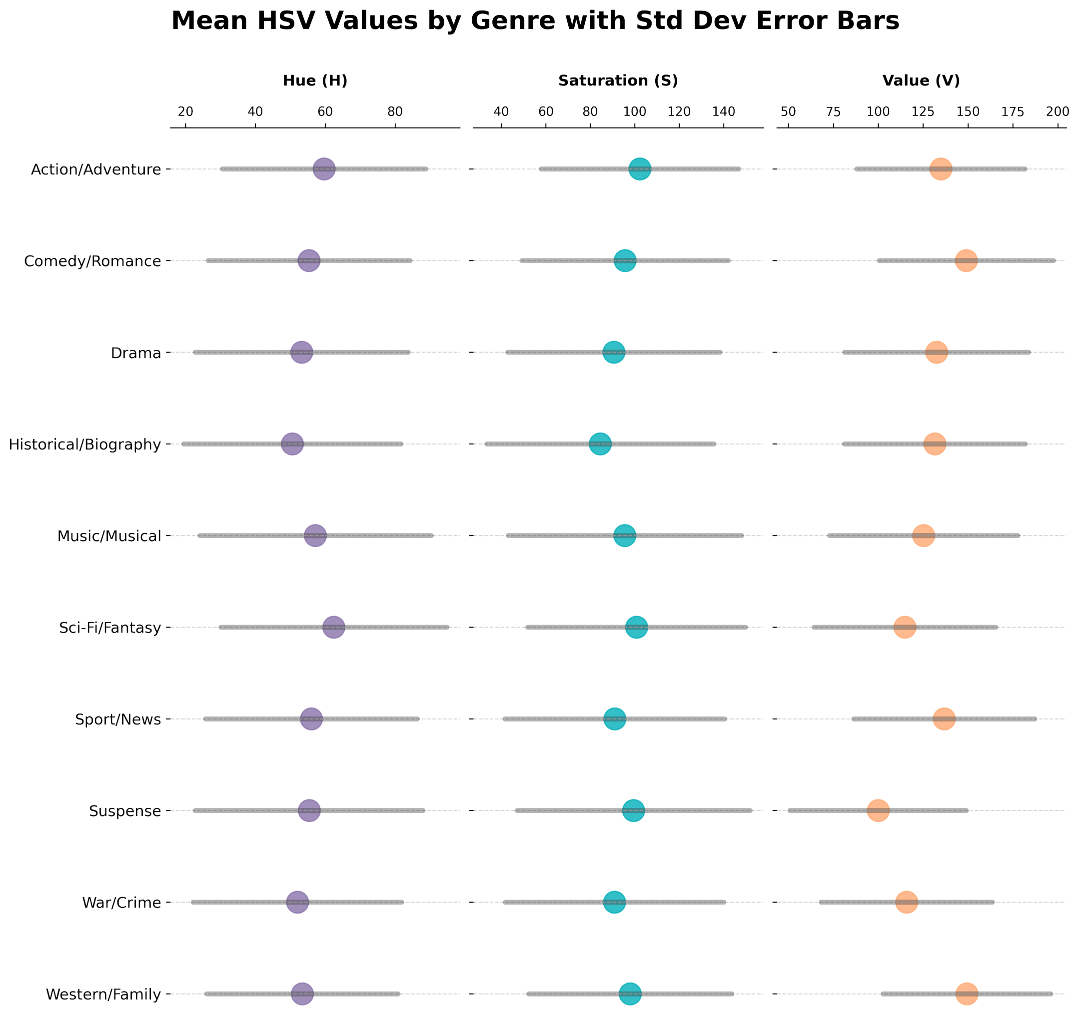

Figure 2: Mean HSV Values by Cha Genre with Standard Deviation Error Bars.

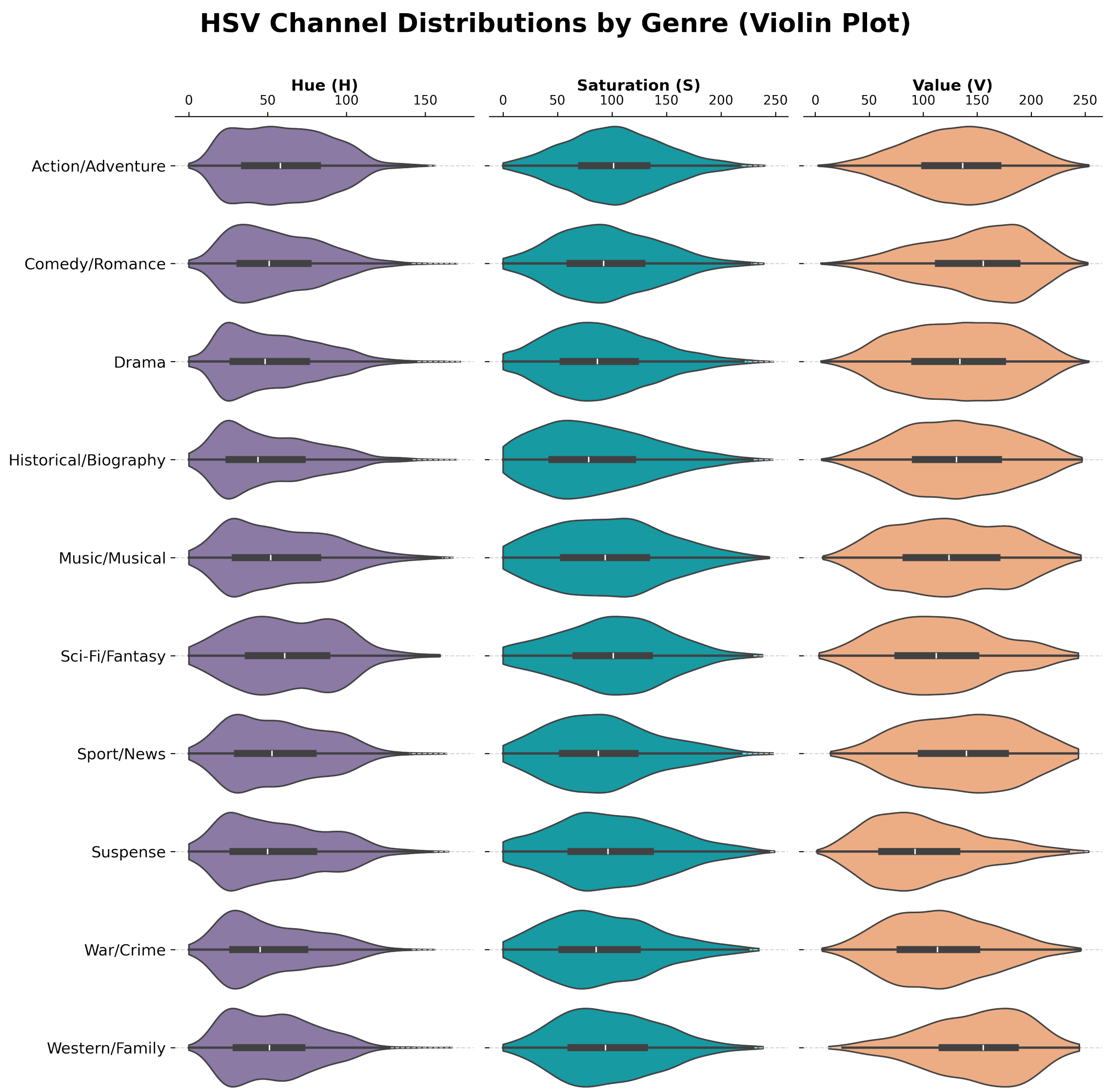

Figure 3: HSV Channel Distributions by Genre (Violin Plot).

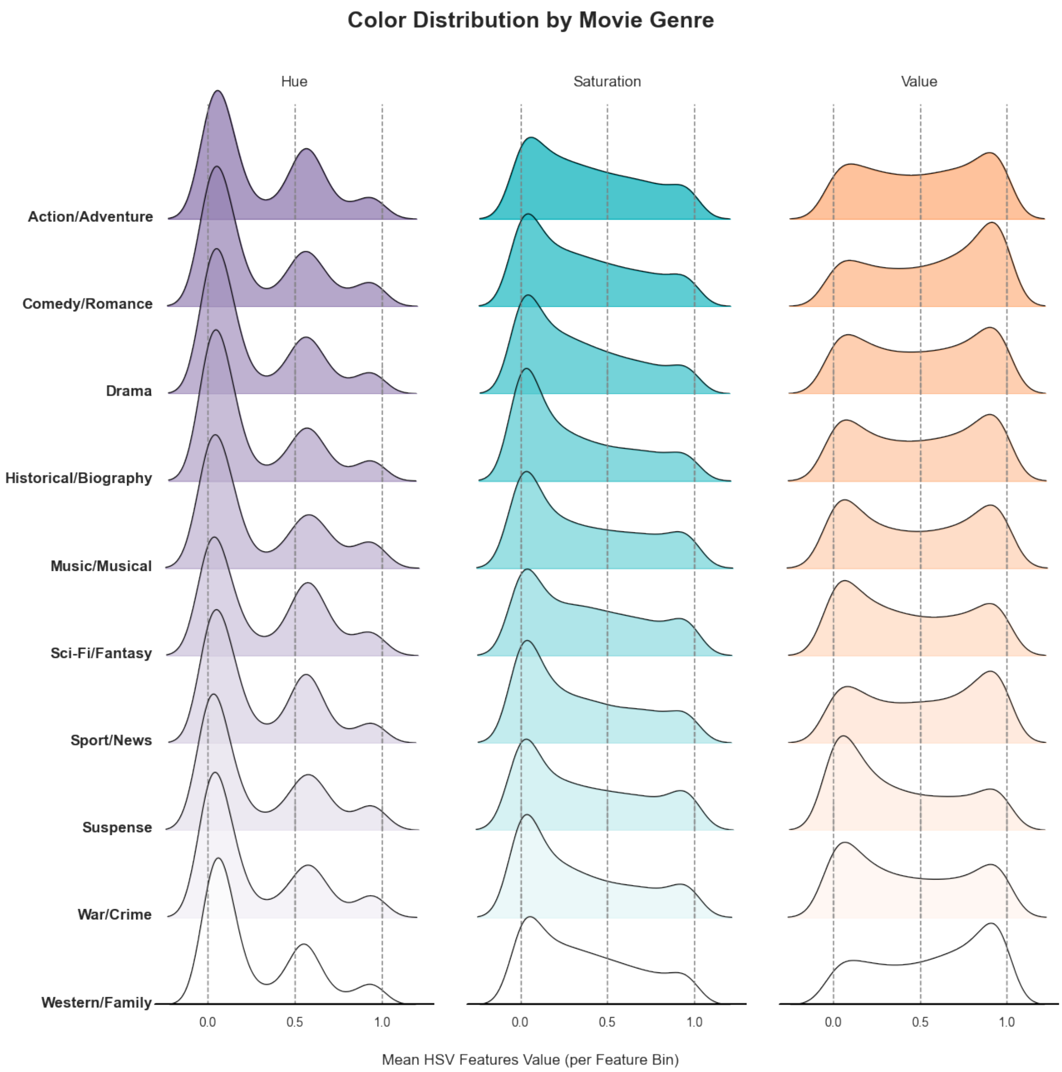

Figure 4: Mean Channel Distribution by Movie Genre (Hue, Saturation, Value).

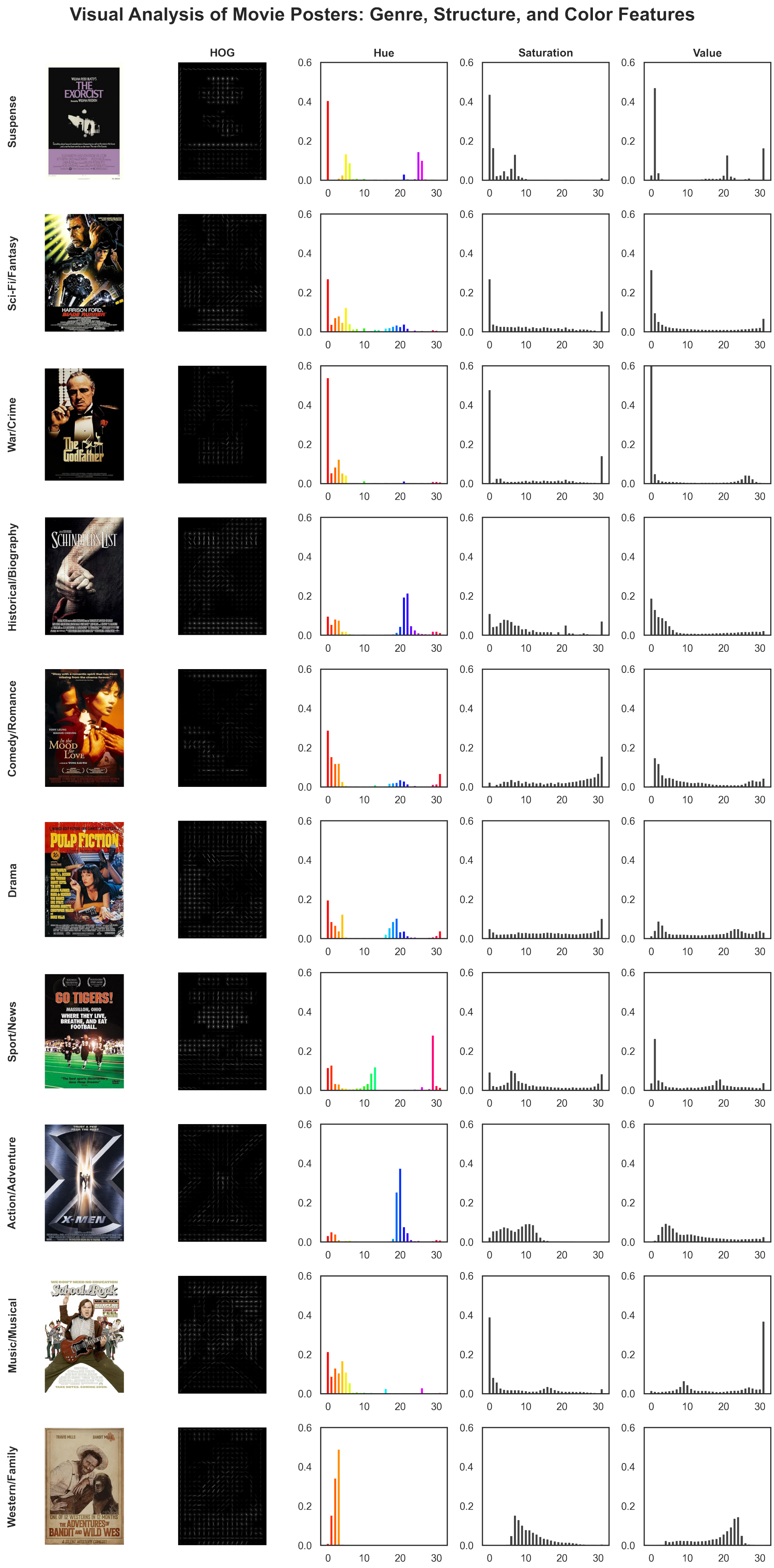

Figure 5: Visual Analysis of Movie Posters: Genre, Structure, and Color Features.

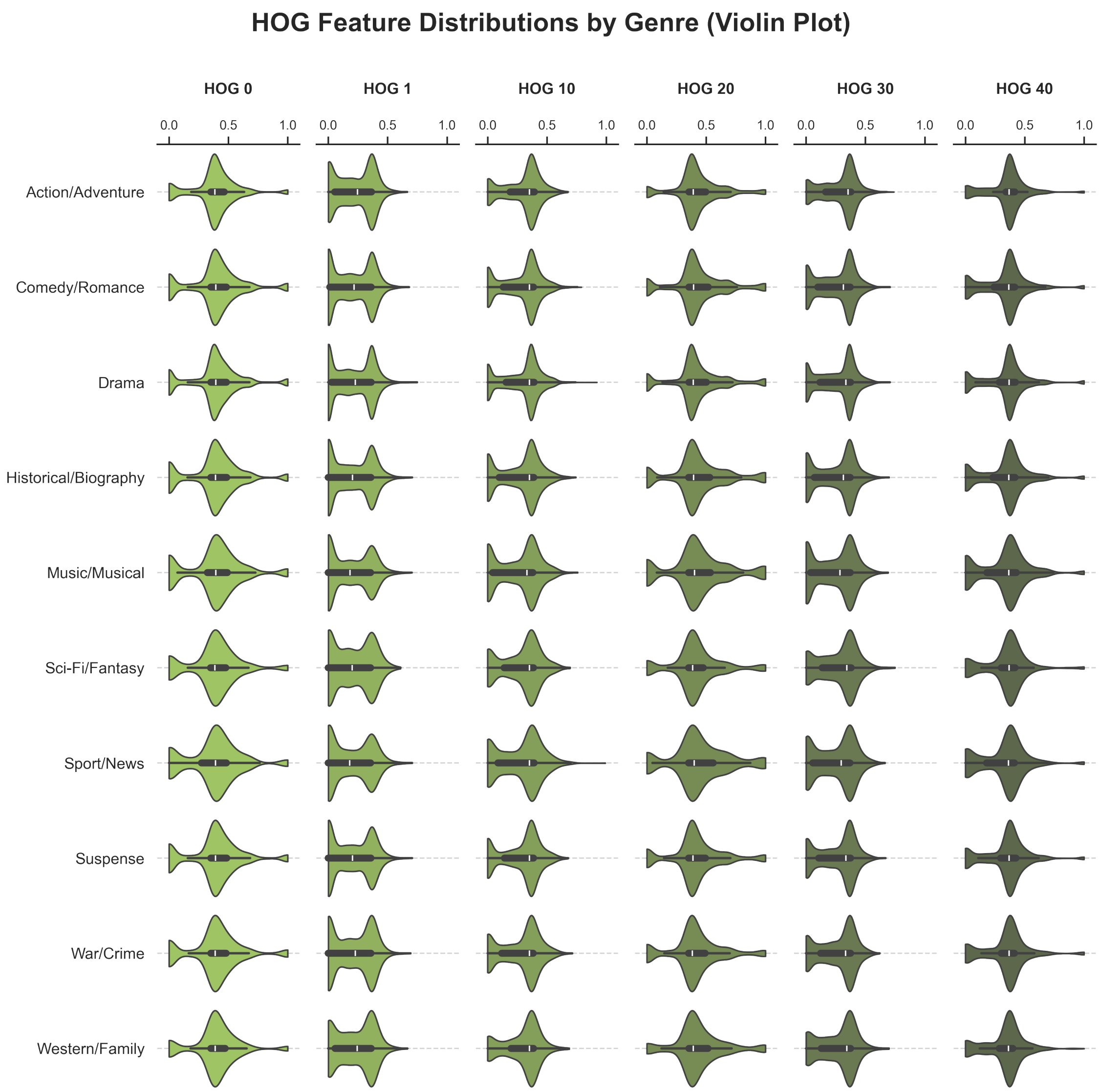

Figure 6: HOG Feature Distributions by Genre (Violin Plot).

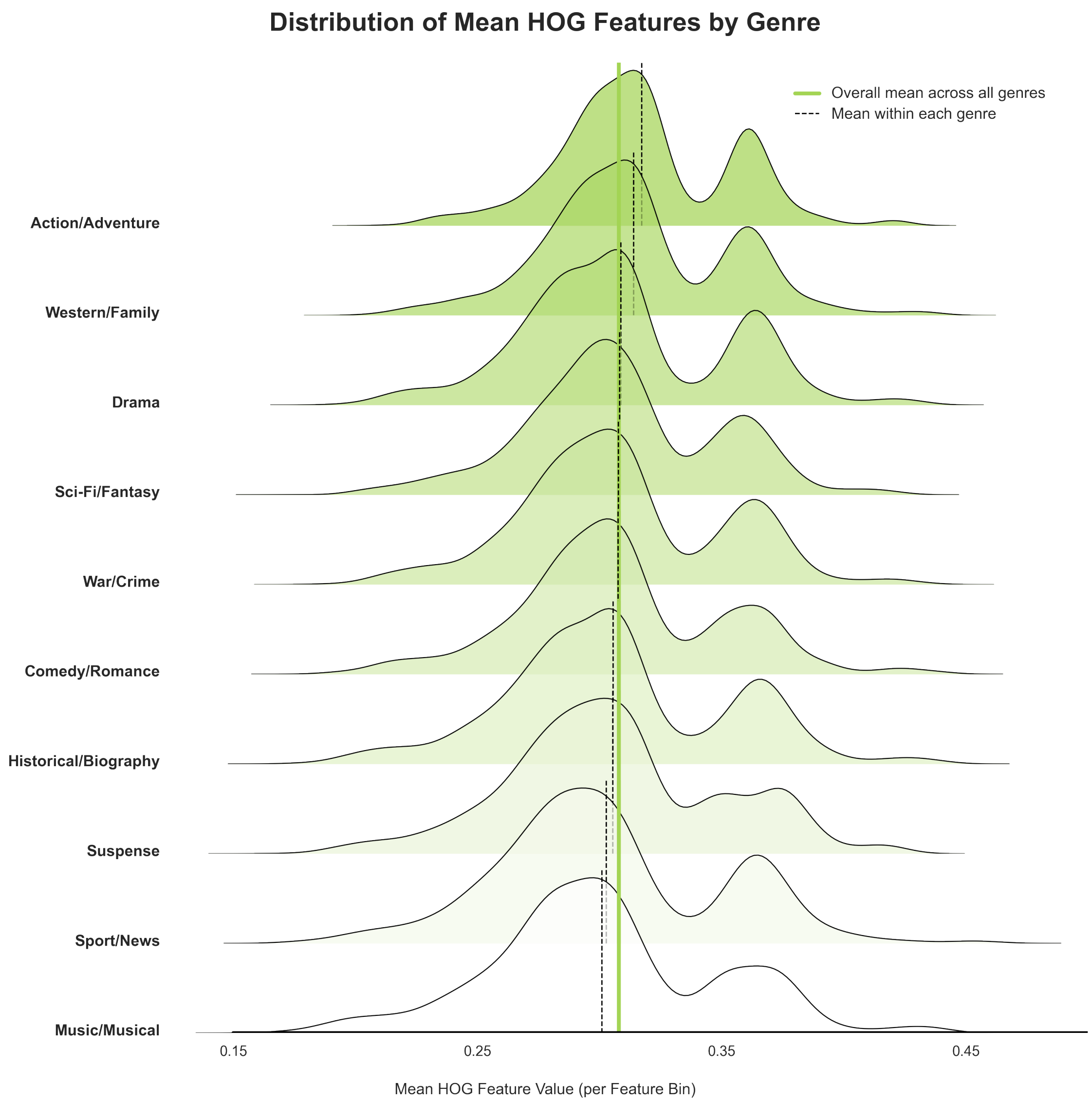

Figure 7: Distribution of Mean HOG Features by Genre.

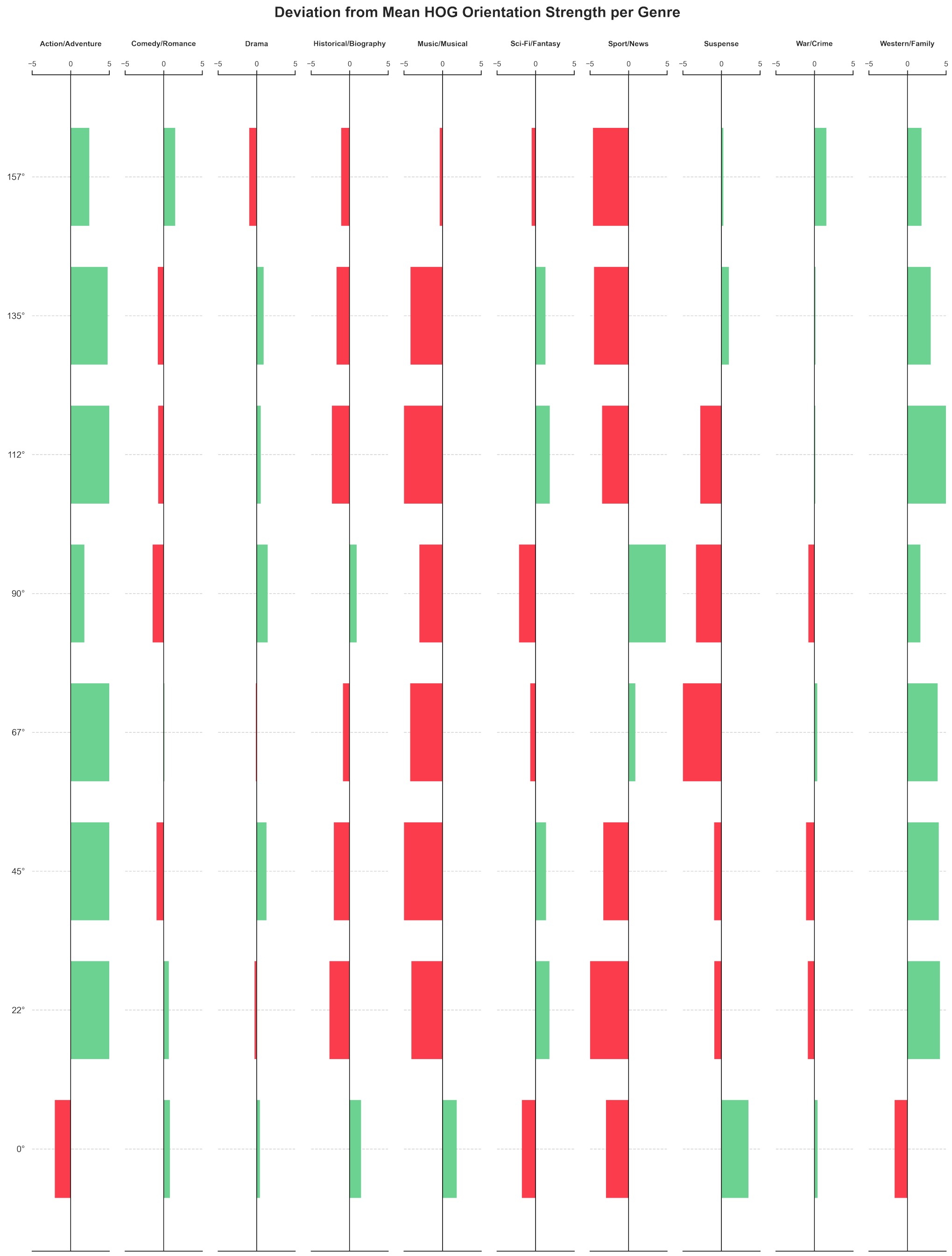

Figure 8: Deviation from Mean HOG Orientation Strength per Genre.

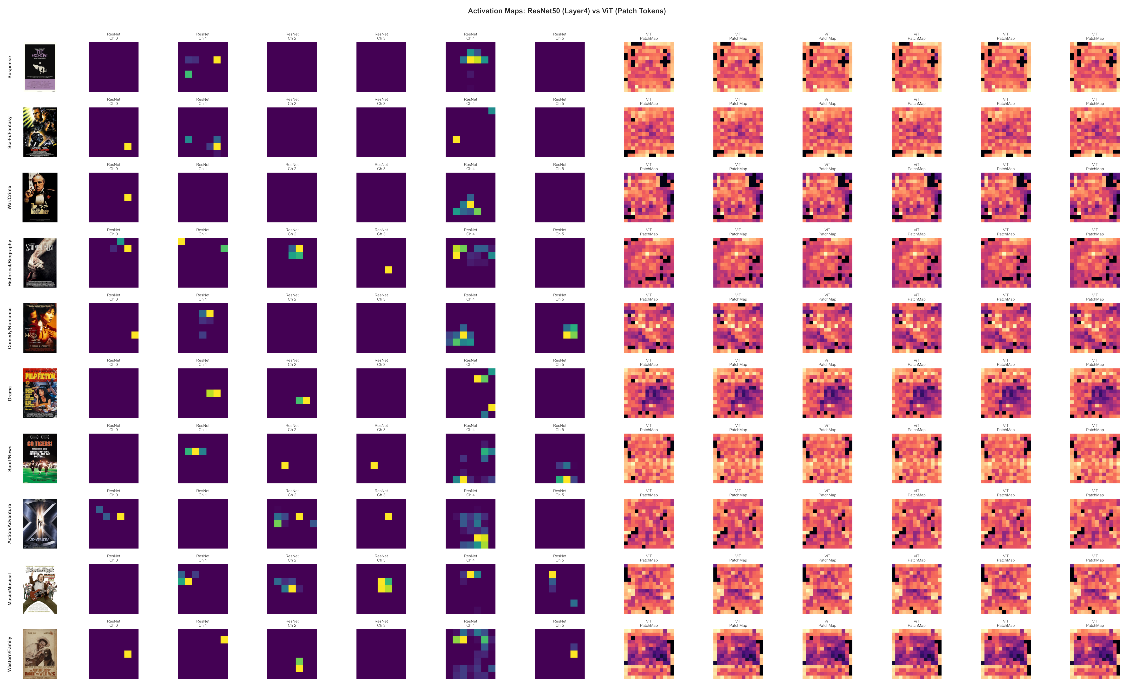

Figure 9: Activation Maps: ResNet50 (Layer4) vs ViT (Patch Tokens).

Figure 10: Side-by-side violin plots of per-genre centroid activation distributions for ResNet50 (left) and ViT (right). Each violin's width encodes the density of average activations across all posters in a genre, with dashed lines indicating the 25th, 50th, and 75th percentiles.

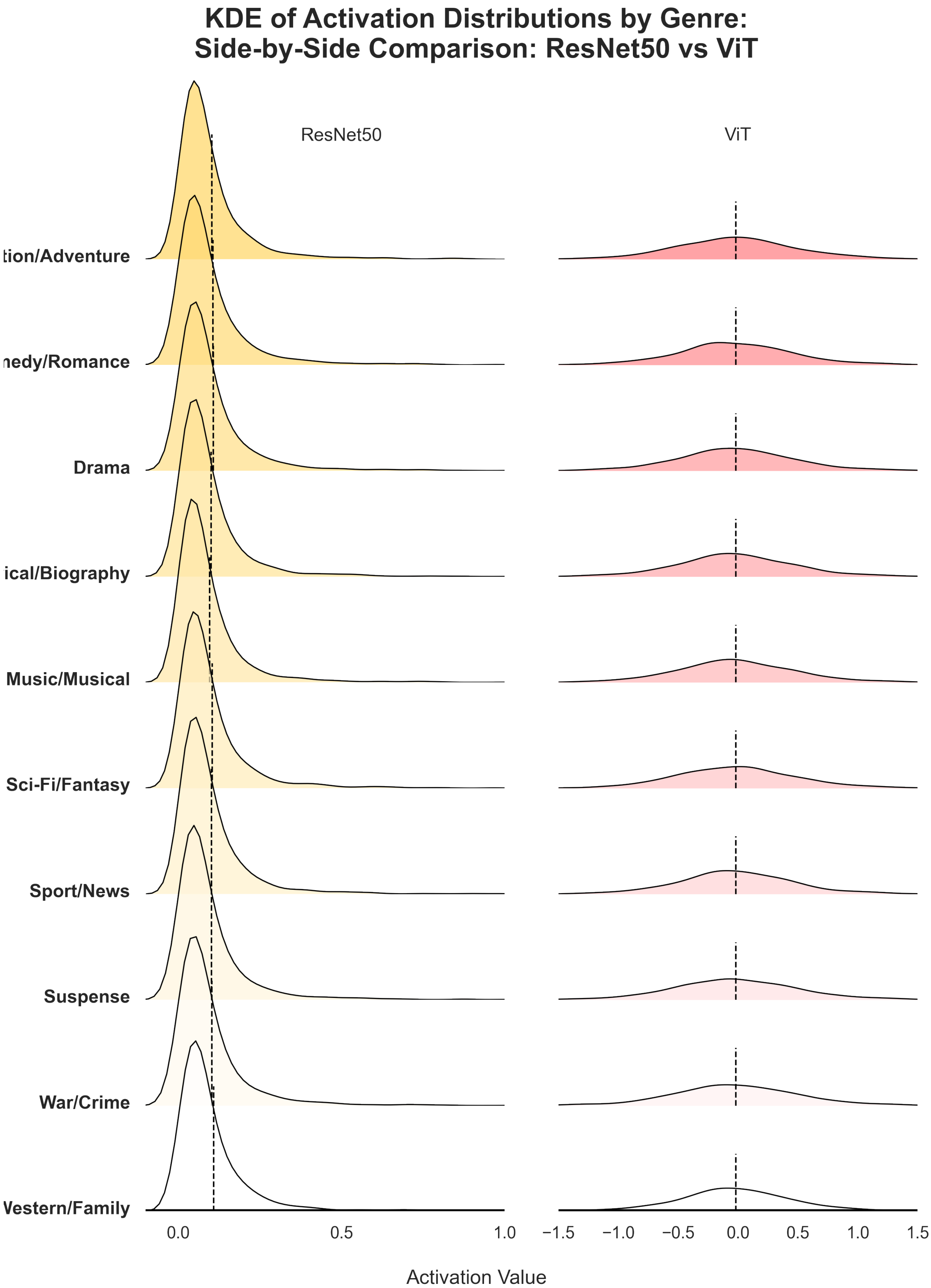

Figure 11: Kernel density estimates of genre-level centroid activations for ResNet50 (left) versus ViT (right). ResNet50 curves show sharp peaks at low activations with long right tails, whereas ViT densities form smoother, symmetric bell shapes centered near zero, highlighting their differing attention patterns.

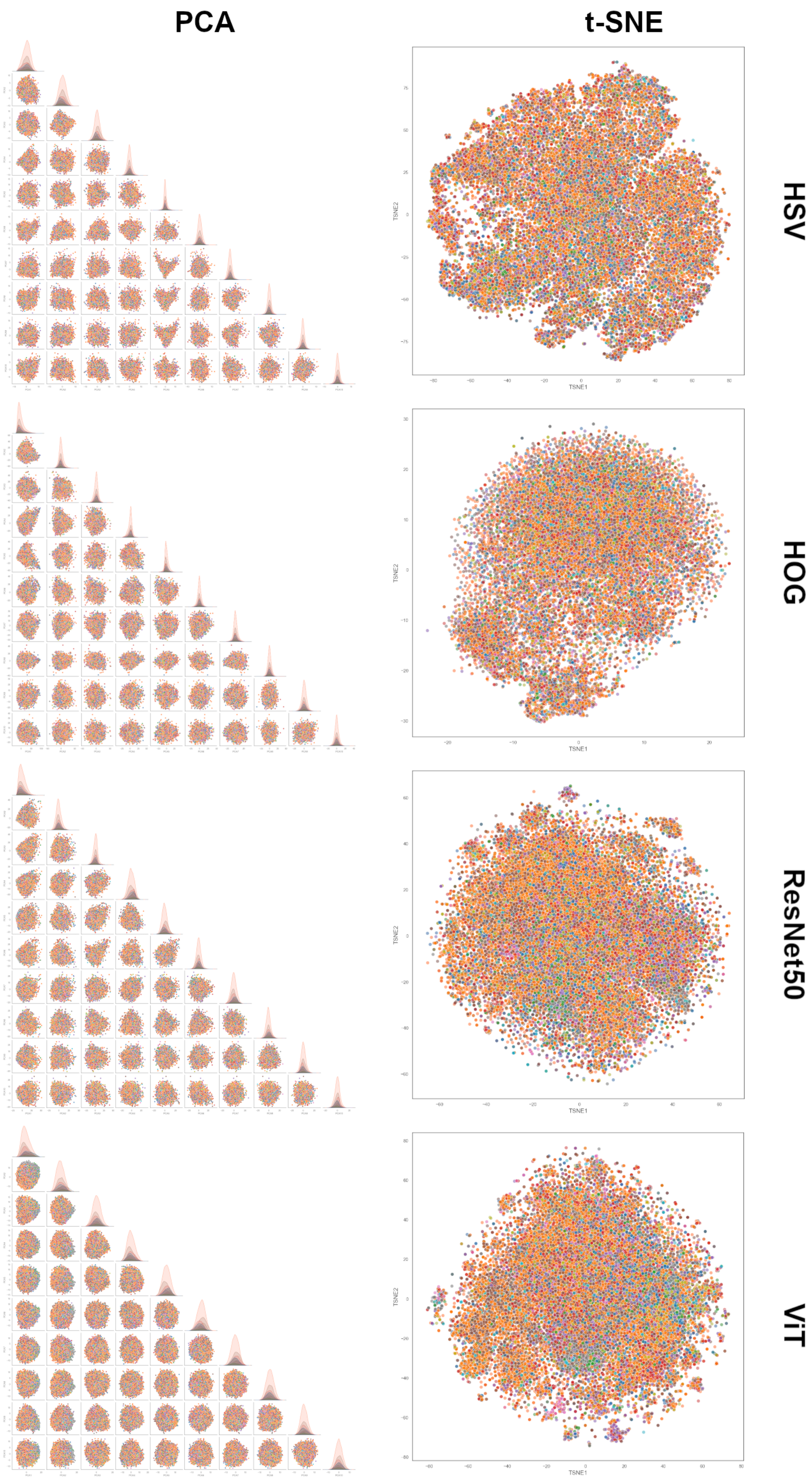

Figure 12: Left: Comparison of pair plots of first 10 PCA components for HSV, HOG, ResNet50, and Vit extracted features. Right: Comparison of t-SNE scatter plots for HSV, HOG, ResNet50, and Vit extracted features

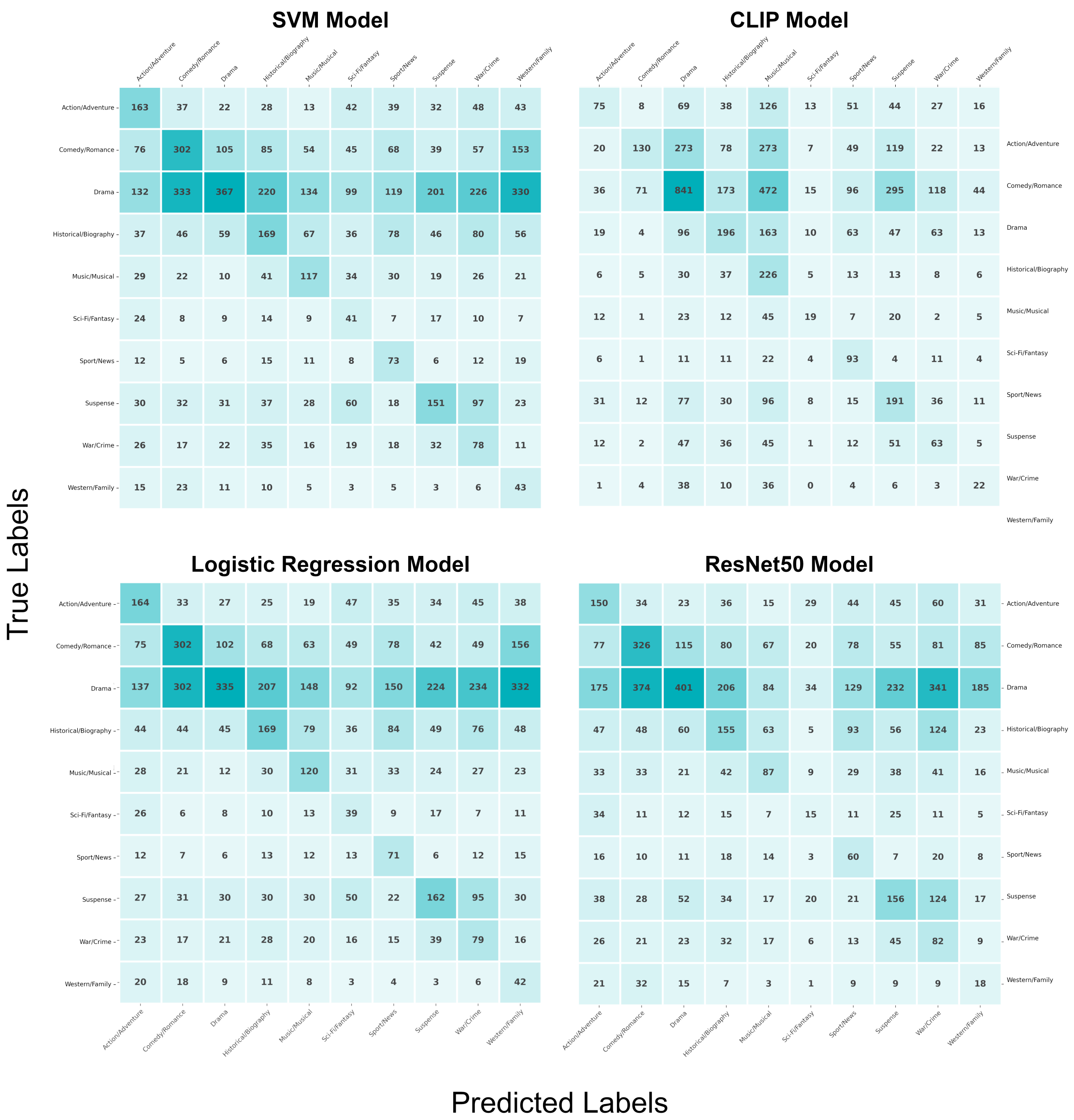

Figure 13: Comparison of confusion matrices for the SVM, Logistic Regression, ResNet50 CNN, and CLIP zero-shot model. Each matrix shows true labels on the vertical axis and predicted labels on the horizontal axis.