Finding Relationships Between Visual Assessment Metrics and User Prompts for Synthetic Generative Images

Steven Broaddus
Computer Science Engineering
Ohio State University
Columbus, Ohio
broaddus.15@buckeyemail.osu.edu

Tommy Tong
Computer Science Engineering
Ohio State University
Columbus, Ohio
tong.414@buckeyemail.osu.edu

Abstract

Prompts serve as the foundation of generative AI outputs, especially in the context of image generation. By altering the language, from changing specific key words to adding more information, the results can vary significantly. In this paper, we focus on how using different key phrases affects the output of generative image models. Specifically, this approach leverages metrics that estimate image quality in an attempt to identify how key phrases impact perceived quality. Using DiffusionDB's dataset of approximately 2,603 prompt-image pairs along with hyperparameter information, we develop predictions models for 49 different image quality metrics, including Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE), Natural Image Quality Evaluator (NIQE), and Perception-Based Image Quality Evaluator (PIQE). Through this analysis, we aim to uncover insights into prompt structure and keyword choice to better understand how natural language influences image generation, offering users greater control over the quality of their generated results and potentially enabling future methods of predicting image quality directly from the prompt, reducing the need for expensive image generation.

1 Introduction

As generative AI continues to grow in popularity, there has emerged an urgent need to streamline methods to filter by quality for any form of content. The areas of the internet that were once feasible to navigate through are quickly becoming flooded with new content, much of which is naturally much lower quality due to the lowered skill floor for content curation and the natural distribution based on computational resource need. As such, benchmarks and new metrics have continued to emerge including new visual quality assessments that also leverage new advances in AI (e.g. Wu et al., 2024). With these new quality metrics, new techniques become feasible for potential optimizations in content generation.

This project aims to systematically investigate how different prompt construction techniques and descriptive keywords-such as those denoting camera models, resolution, scene themes, and artistic styles-influence the output of generative image models. It is well known that prompt engineering is an important staple for efficient generative AI usage, and this project aims to further the understanding of how it can be utilized by the user to optimize standardized quality metrics.

2 Datasets

There exist several large-scale datasets that pair prompts with generated images, including diffusiondb (Wang et al., 2022), open-prompts, and Lexica. These datasets provide a foundation for our analyses, with millions of prompt-image pairs. Prior work in this space has focused directly on prompt behaviors Jahani et al.'s (2024), Liu and Chilton's (2022), Shin et al.'s (2024), automated prompt systhesis Cao et al.'s (2023), prompt-induced bias Shin et al.'s (2024), and seantic alignment between prompt and output Zhan et al.'s (2024). However, few studies have applied systematic computer vision analyses that utilize vision quality assessment metrics to quantitatively evaluate prompt effects across model outputs.

To address this gap, our approach involves: (1) identifying key visual quality metrics; (2) building a dataset-agnostic pipeline for both image and prompt analysis; (3) characterizing prompt structures and keywords; and (4) comparing trends across datasets and domains. Through this analysis we aim to surface actionable insights that help users craft more effective prompts for specific metrics.

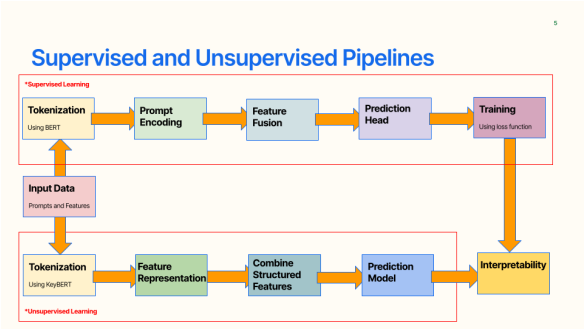

Figure 1: Pipelines for data processing.

3 Metrics

Using Chen's (2024) PyTorch Toolbox for Image Quality Assessment, we were able to generate the mainstream full reference and no reference metrics. In this section, we will elaborate on 8 of the most popular metrics out of the 49 generated. These will be Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE), Natural Image Quality Evaluator (NIQE), Perception-based Image Quality Evaluator (PIQE), Deep Bilinear Convolutional Neural Network (DBCNN), Multi-scale Image Quality Transformer (MUSIQ), Perceptual Assessment of Quality for Quality Prediction (PAQ2PIQ), Neural Image Assessment (NIMA), and No-Reference Quality Metric (NRQM). These metrics will help define and identify the prompt structures and keywords to focus on.

4 Experimental Methodology

We evaluate two parallel modeling pipelines to predict image quality metrics based solely on the textual prompts used to generate the images. The supervised pipeline embeds prompts into continuous representations using BERT [CLS] token embeddings, supplemented with handcrafted prompt features, and trains supervised regressors to predict quality scores. In contrast, the unsupervised pipeline extracts keyphrases from prompts using KeyBERT, represents them through TF-IDF vectorization, and applies simpler regressors without relying on deep embeddings. Both pipelines employ cross-validation and are evaluated using mean squared error (MSE), mean absolute error (MAE), and coefficient of determination (R^2) as performance metrics. Please refer to Figure 1 for an overview of the two different pipelines.

4.1 Pre Processing

For the supervised pipelines, each prompt is embedded into a 768-dimensional vector using the [CLS] token from a pre-trained BERT model (`bert-base-uncased`). Additional handcrafted features are extracted from the prompts, including word count, unique word count, comma count, a custom complexity score, number of adjectives, and number of named entities. These structured features are standardized using a `StandardScaler`, and then concatenated with the BERT embeddings to form the final input feature set for supervised models.

For the unsupervised pipelines, KeyBERT is used to extract a set of top keywords for each prompt, using a sentence-transformer model (`all-MiniLM-L6-v2`). These extracted keyphrases are cached to disk to improve efficiency and reproducibility. The keyphrases are then vectorized using TF-IDF, treating multi-word phrases as atomic tokens by customizing the tokenizer. The resulting sparse TF-IDF matrix is used directly as input to the unsupervised models without further scaling or normalization.

4.2 Processing

In the supervised pipeline, the feature matrix is constructed by concatenating the 768-dimensional BERT [CLS] embeddings with standardized prompt-derived features. Models including multilayer perceptrons (MLP), random forests, gradient boosting regressors, and support vector regressors are trained using five-fold cross-validation. Hyperparameters are tuned via grid search over a predefined space for each model class. In the unsupervised pipeline, prompts are first transformed into cached KeyBERT keyword sets, which are then vectorized using TF-IDF. Regressors are trained directly on these sparse TF-IDF features following the same cross-validation and grid search procedure.

4.3 Post Processing

After model training and evaluation, SHAP (SHapley Additive exPlanations) analysis is conducted to interpret model predictions. For supervised models, SHAP values are computed on the combined BERT and structured features, while for unsupervised models, SHAP values are computed over the TF-IDF feature space. The mean absolute SHAP value for each feature is aggregated and saved for anal-

IQA Metric	Description
BRISQUE	No-reference IQA metric that evaluates image distortion and naturalness.
NIQE	No-reference IQA metric that estimates image quality based on statistical properties
PIQE	No-reference IQA metric that estimates perceptual quality based on spatial and statistical features
DBCNN	Deep-learning model trained on large-scale datasets
MUSIQ	Transformer-based IQA metric that analyzes images at multiple scales
PAQ2PIQ	Deep-learning model predicting perceptual image quality from human-labeled datasets
NIMA	A model that predicts human aesthetic ratings based on deep-learrning techniques
NRQM	No-reference IQA metric that uses natural scene statistics to predict perceived quality

Table 1: 8 Most Common Image Quality Assessments (IQA) and Descriptions

ysis. In addition, summary plots and per-feature CSV files are generated to facilitate further interpretation of key features driving model performance across both pipelines.

5 Results

5.1 Model Performance Metrics

Model Evaluation Metrics. We evaluated model performance using the coefficient of determination (R^2), mean absolute error (MAE), and mean squared error (MSE). These metrics offer complementary perspectives: R^2 measures the proportion of variance explained by the model, while MAE and MSE reflect the magnitude of prediction errors, with MSE penalizing large errors more heavily. Together, they provide a robust view of predictive accuracy and stability across both supervised and unsupervised approaches.

Supervised vs. Unsupervised Models. As shown in Figure 2, supervised models consistently outperform unsupervised ones across most metrics, with notably higher R^2 values and lower MAE/MSE values (Figures 3 and 4). Among supervised models, **Random Forest** and **Gradient Boosting** generally achieve the highest R^2 scores and lowest errors, suggesting strong generalization capabilities with modest complexity. **SVR** also performs well but tends to have slightly higher variance in performance. On the other hand, **ANN** (MLPRegressor) underperforms relative to the other supervised models, especially in terms of R^2, potentially due to overfitting or sensitivity to initialization and learning rate in smaller datasets.

Among the unsupervised models, **Gradient Boosting** and **Random Forest** using TF-IDF keyword features demonstrate relatively stronger performance, although still significantly behind their supervised counterparts. **KeyBERT+MLP**, while intuitive and interpretable, frequently achieves the lowest R^2 scores and highest error rates, reflecting the limitations of sparse keyword representations in capturing the rich semantics of prompts.

Understanding Low R^2 Scores. Although R^2 values across models rarely exceed 0.5, this should not be interpreted as a failure of the approach. Instead, it reflects the inherent noise and unpredictability in the prompt-to-image generation process. Generative image models often include elements of randomness, multimodality, and latent conditioning (e.g., training biases, aesthetic priors) that are not captured by the prompt alone. Consequently, even highly descriptive prompts may not deterministically map to the final image quality, introducing irreducible noise into the target signal. Despite this, the models' ability to explain even 20–40% of the variance across multiple image quality metrics indicates that prompts contain genuine, learnable structure related to output quality. These results highlight the potential for prompt-based analytics in downstream tasks like prompt optimization or filtering without requiring full image generation.

Implications for Model Design. These findings suggest that dense, contextualized representations like BERT embeddings paired with structured prompt features offer superior predictive value for modeling image quality from prompts. Models leveraging these inputs outperform traditional keyword-based pipelines by a substantial margin. Random Forest and Gradient Boosting emerge as robust, high-performing choices for supervised regression in this domain, striking a balance between interpretability and predictive power.

5.2 SHAP Analyses

SHAP Analysis of Prompt Feature Contributions. To investigate the relationship between

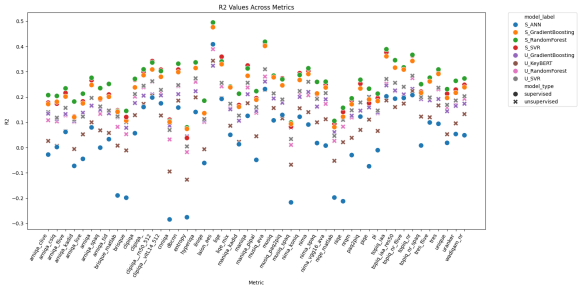

Figure 2: R^2 scores across all models and metrics. Higher is better. Supervised Random Forest and Gradient Boosting models consistently outperform others.

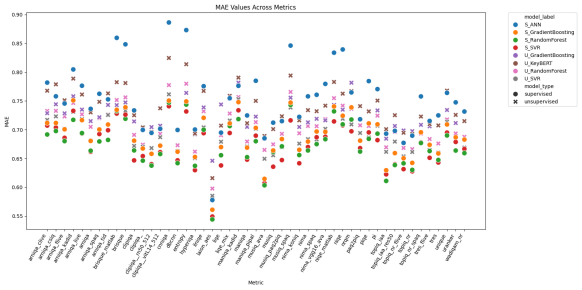

Figure 3: MAE scores across models and metrics. Lower is better. Supervised models generally yield smaller prediction errors.

prompt composition and resulting image quality, a SHAP (SHapley Additive exPlanations) analysis was conducted to quantify the influence of individual prompt features on various quality metrics. Figure 5 illustrates an example SHAP explanation for the BRISQUE metric using a Random Forest model. Notably, tokens such as *portrait*, *painting*, and *photography* exhibit strong feature contributions, indicating a substantial effect on the predicted BRISQUE values. Given that BRISQUE measures deviations from natural scene statistics (NSS) in the spatial domain, lower scores correspond to fewer distortions, a more natural appearance, and higher perceptual quality. Thus, prompt features with strong negative SHAP values are associated with improved image outputs. Figure 6 presents a cumulative SHAP contribution curve summarizing the overall influence of key phrases across all metrics. The x-axis represents ranked prompt features, ordered by their individual SHAP impact, while the y-axis reflects the cumulative percentage of the model's total explainable output variance. The curve indicates that approximately 200 key phrases are required to account for 80% of the cumulative SHAP contribution, highlighting the wide distribution of influential language

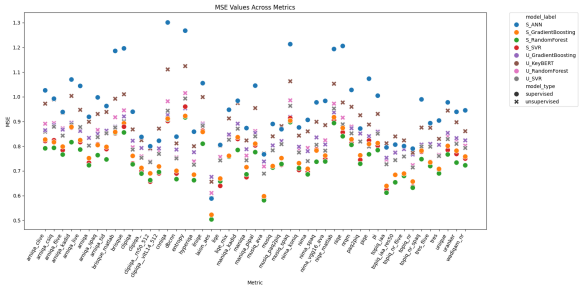

Figure 4: MSE scores across models and metrics. Higher is better. Gradient Boosting and Random Forest maintain low error magnitudes across metrics.

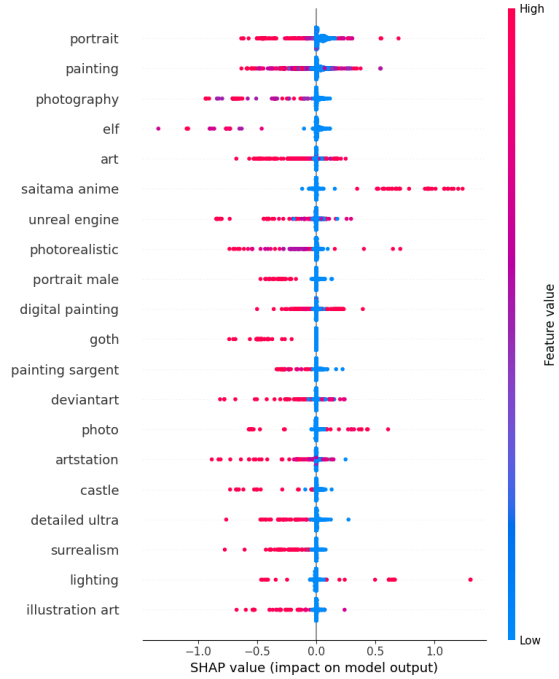

Figure 5: Example SHAP explanation for the Random Forest model for the BRISQUE metric (lower values represent a better BRISQUE metric)

components within prompts.

5.3 Important Keyphrases

Top tokens and Word Cloud Visualizations The results demonstrate that specific tokens and phrases significantly influence the characteristics of generated images. Figure 7 identifies tokens with the strongest impact. Among these, the three most impactful words—*film*, *painting*, and *portrait*—exhibit a pronounced effect on image generation. A word cloud visualization, shown in Figure 8, further highlights these influential tokens where size equates to impact.

Cluster Analysis Figure 9 presents a UMAP-based cluster visualization annotated with thematic

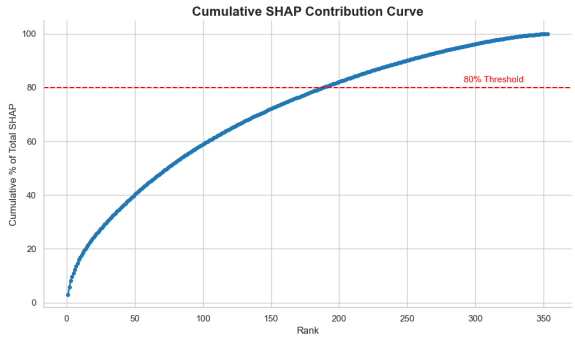

Figure 6: Cumulative SHAP Contribution Curve

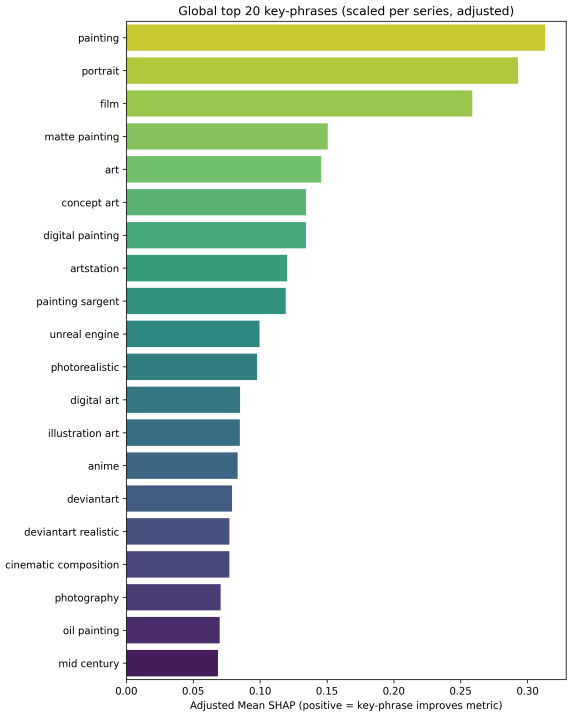

Figure 7: Top 20 Keyphrases across all 4 unsupervised models for all 49 metrics for all 2603 image-prompt pairs.

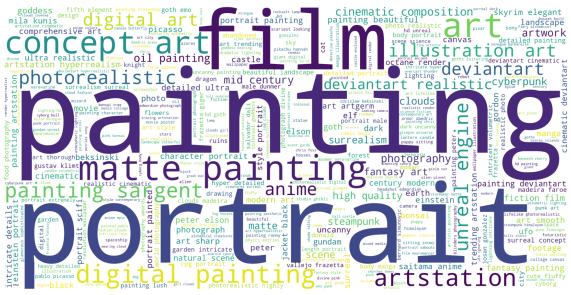

Figure 8: Word Cloud Representation of Top Keyphrases

labels. Densely packed and centralized clusters correspond to distinct aesthetic and thematic categories, such as *Artistic Influences* and *Dark Surrealism*. The size of each cluster reflects thematic variability: larger clusters, such as *Artistic Influences*, encompass a broader range of variations, whereas smaller clusters, such as *Divine Cloud Imagery*, produce more consistent and homogeneous outputs. The spatial arrangement of clusters further reveals semantic relationships. For instance, *Artistic Influences* and *Divine Cloud Imagery* are positioned furthest apart, indicating significant differences in both visual and semantic style. In contrast, clusters positioned closely together, such as *Science Fiction Scenes* and *Alien Encounters*, exhibit overlapping thematic qualities. The clear separation between clusters suggests that the selected features effectively capture meaningful distinctions in the data.

6 Conclusion

This paper investigates the influence of specific prompt keywords on the output of generative image models. By analyzing various tokens and keyphrases, terms such as *painting*, *portrait*, and *film* significantly impact the visual characteristics of generated images. Through the use of different IQA metrics, models, and performing SHAP analysis, we have provided insights into the relationship between prompt structure and image quality.

The results show that prompt engineering plays a crucial role in shaping the thematic and aesthetic qualities of generated content. In particular, Figure 9 reveals how distinct prompt features correspond to different thematic categories, with clear separations between clusters reflecting meaningful distinctions in both visual and semantic aspects.

Our findings highlight the potential of prompt optimization to enhance user control over the generated image quality. Further research could focus on refining prompt structures for more fine-grained control, as well as exploring the integration of predictive models for image quality directly from the prompt, which would reduce the need for resource-intensive image generation processes. Overall, this work contributes to a deeper understanding of how prompt engineering can be utilized to optimize the performance of generative image models.

Acknowledgments

We would like to thank our course instructor, Professor Kumar, for their guidance throughout the duration of this project. Their insights and feedback were instrumental in shaping the direction of

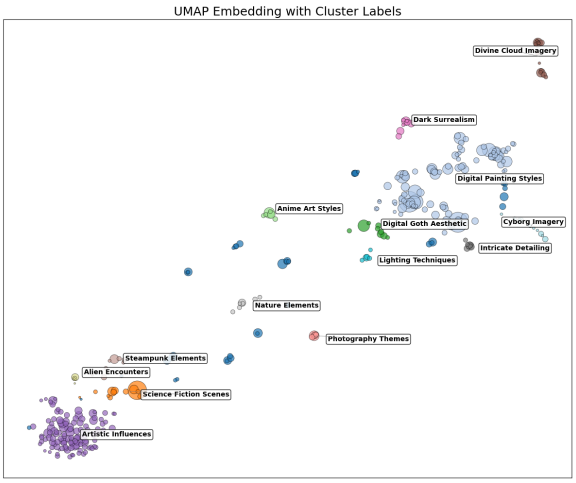

Figure 9: UMAP Embedding with Clustering Labels

our research.

References

Tingfeng Cao, Chengyu Wang, Bingyan Liu, Ziheng Wu, Jinhui Zhu, and Jun Huang. 2023. Beautifulprompt: Towards automatic prompt engineering for text-to-image synthesis. *Preprint*, arXiv:2311.06752.

Chao-Feng Chen. 2024. IQA-PyTorch: A pytorch implementation for image quality assessment. https://github.com/chaofengc/IQA-PyTorch. Accessed: 2025-04-25.

Eaman Jahani, Benjamin S. Manning, Joe Zhang, Hong-Yi TuYe, Mohammed Alsobay, Christos Nicolaides, Siddharth Suri, and David Holtz. 2024. As generative models improve, we must adapt our prompts. *Preprint*, arXiv:2407.14333.

Vivian Liu and Lydia B Chilton. 2022. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Philip Wootaek Shin, Jihyun Janice Ahn, Wenpeng Yin, Jack Sampson, and Vijaykrishnan Narayanan. 2024. Can Prompt Modifiers Control Bias? A Comparative Analysis of Text-to-Image Generative Models. *Preprint*, arXiv:2406.05602.

Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. 2022. Large-scale prompt gallery dataset for text-to-image generative models. *arXiv:2210.14896 [cs]*.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Chunyi Li, Liang Liao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtai Zhai, and Weisi Lin. 2024. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *International Conference on Machine Learning (ICML)*. Equal Contribution by Wu, Haoning and Zhang, Zicheng. Project Lead by Wu, Haoning. Corresponding Authors: Zhai, Guangtai and Lin, Weisi.

Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. 2024. A General Protocol to Probe Large Vision Models for 3D Physical Understanding. *Preprint*, arXiv:2310.06836.

A Appendix

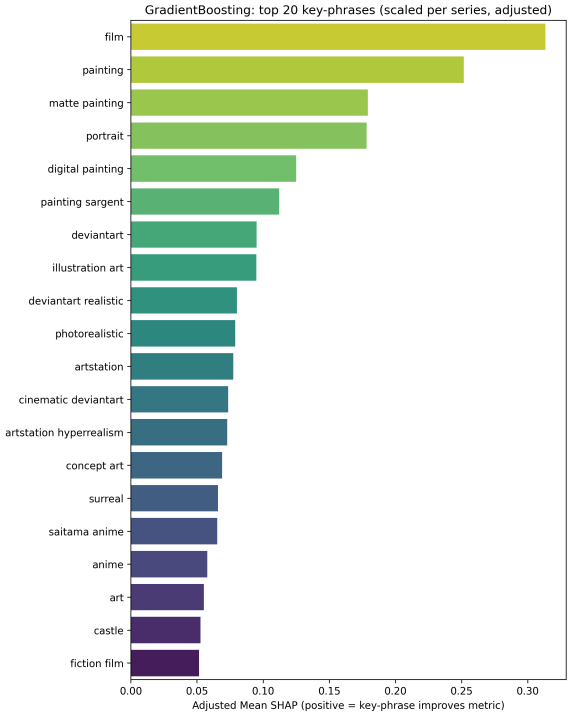

Figure 10: Top 20 Key-Phrases for GradientBoosting

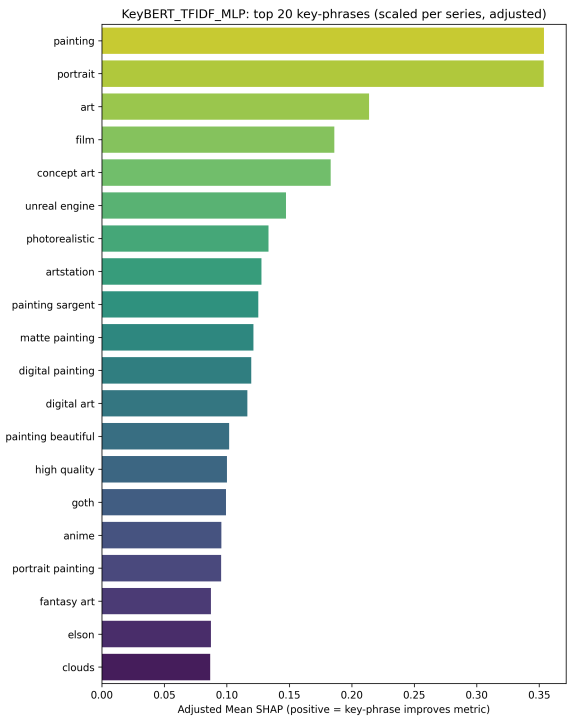

Figure 11: Top 20 Key-Phrases for KeyBert

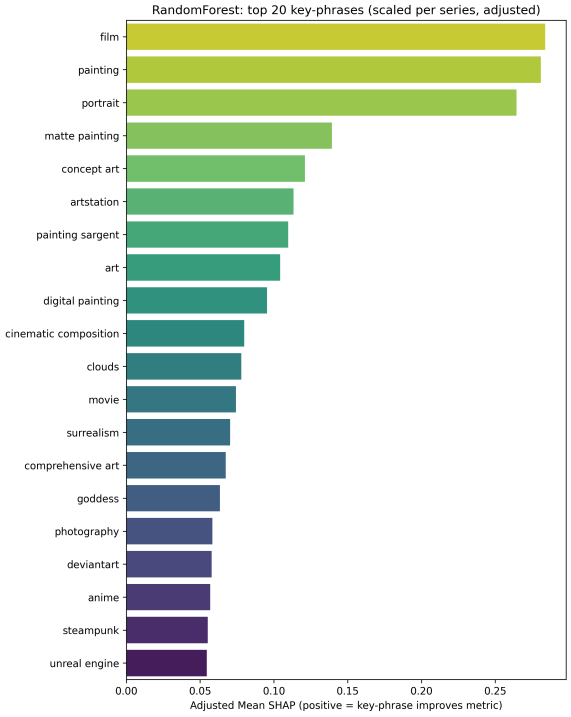

Figure 12: Top 20 Key-Phrases for RandomForest

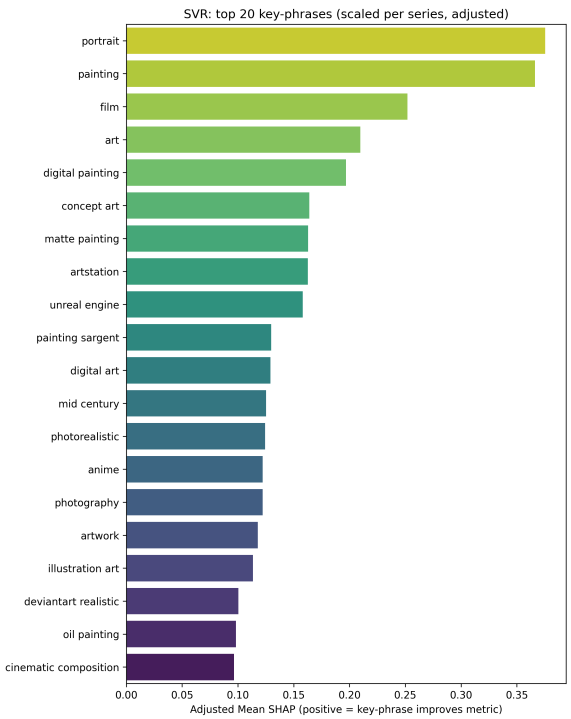

Figure 13: Top 20 Key-Phrases for SVR

IQA Metric	Description
Arniqa	No-reference IQA metric assessing image distortion and quality.
Arniqa_clive	Arniqa variant optimized for the CLIVE dataset.
Arniqa_csiq	Arniqa model trained on CSIQ dataset, focusing on perceptual quality.
Arniqa_flive	Arniqa trained on the LIVE dataset for no-reference quality prediction.
Arniqa_kadid	Arniqa metric for Kadid dataset with natural image quality characteristics.
Arniqa_live	Live dataset-based model for no-reference quality assessment.
Arniqa_spaq	Arniqa variant trained on SPAQ dataset for quality prediction.
Arniqa_tid	Arniqa model designed for the TID dataset, predicting perceived quality.
Brisque	No-reference IQA metric for evaluating image distortion and naturalness.
Brisque_matlab	MATLAB implementation of BRISQUE for quality assessment.
Clipiqa	Deep learning-based IQA metric using CLIP model for perceptual quality.
Clipiqa_	Variation of Clipiqa using alternative preprocessing for quality prediction.
Clipiqa__rn50_512	Clipiqa variant using ResNet50 for image quality evaluation.
Clipiqa__vitL14_512	Clipiqa model utilizing Vision Transformer for perceptual quality.
Cnniqa	CNN-based IQA metric for no-reference quality evaluation.
Dbcnn	Deep learning-based model trained on large-scale datasets for IQA.
Entropy	Entropy-based metric assessing randomness in an image for quality.
Hyperiqa	Hyperparameters-based deep learning model for image quality assessment.
Ilniqe	No-reference IQA metric using local image features for quality evaluation.
Laion_aes	Aesthetic score model that evaluates images for quality.
Liqe	No-reference IQA metric based on statistical image features.
Liqe_mix	Hybrid model combining multiple feature sets for image quality prediction.
Maniqa	Deep learning-based model for general image quality assessment.
Maniqa_kadid	ManIQA optimized for Kadid dataset, focused on perceptual quality.
Maniqa_pipal	ManIQA variant for PIPAL dataset, evaluating perceptual quality.
Musiq	Transformer-based IQA model that analyzes image quality at multiple scales.
Musiq_ava	AVA-based dataset model for aesthetic quality evaluation.
Musiq_paq2piq	MUSIQ model trained on PAQ2PIQ dataset for perceptual quality prediction.
Musiq_spaq	MUSIQ trained on SPAQ dataset for perceptual image quality.
Nima	Deep learning model predicting human aesthetic ratings for images.
Nima_koniq	NIMA model trained on KONIQ dataset for aesthetic image quality prediction.
Nima_spaq	NIMA trained on SPAQ dataset for perceptual quality evaluation.
Nima_vgg16_ava	NIMA variant using VGG16 features for aesthetic score prediction.
Niqe	No-reference IQA metric based on natural scene statistics for quality.
Niqe_matlab	MATLAB implementation of NIQE for quality assessment.
Nrqm	No-reference IQA metric using natural scene statistics for image quality.
Paq2piq	Deep learning model predicting perceptual quality from human-labeled datasets.
Pi	Perceptual IQA model based on image feature statistics.
Piqe	Image quality evaluation metric focused on perceptual features.
Qalign	Metric for assessing alignment and similarity in image quality.
Qalign_8bit	Version of Qalign for evaluating 8-bit image quality.
Topiq_iaa	Top-IQA model using IAA for quality prediction.
Topiq_iaa_res50	Top-IQA with ResNet50 architecture for image quality prediction.
Topiq_nr	No-reference Top-IQA model for image quality evaluation.
Topiq_nr_flive	Top-IQA model trained on the FLIVE dataset for no-reference quality.
Topiq_nr_spaq	No-reference Top-IQA model trained on the SPAQ dataset.
Tres	Model for quality assessment based on texture features in images.
Tres_flive	Variation of Tres model focused on the FLIVE dataset for quality prediction.
Unique	A unique model evaluating image quality based on specific features.
Uranker	Ranking-based model for evaluating image quality.
Wadiqam_nr	Wadiqam for no-reference quality prediction with image-specific focus.

Table 2: All Calculated Image Quality Assessment (IQA) Metrics and Descriptions

Keyphrase	Adjusted Scaled Mean SHAP
painting	0.3132
portrait	0.2929
film	0.2588
matte painting	0.1507
art	0.1457
concept art	0.1343
digital painting	0.1342
artstation	0.1203
painting sargent	0.1192
unreal engine	0.0995
photorealistic	0.0977
digital art	0.0851
illustration art	0.0849
anime	0.0832
deviantart	0.0790
deviantart realistic	0.0771
cinematic composition	0.0770
photography	0.0705
oil painting	0.0699
mid century	0.0685
artwork	0.0678
cyberpunk	0.0650
clouds	0.0635
scene	0.0617
matte	0.0598
surrealism	0.0598
fantasy art	0.0593
steampunk	0.0592
high quality	0.0581
artstation hyperrealism	0.0581
painting beautiful	0.0573
surreal	0.0566
fiction film	0.0565
goddess	0.0563
movie	0.0545
comprehensive art	0.0530
character portrait	0.0526
peter	0.0525
picasso	0.0524
beksinski	0.0519
gordon	0.0518
cinematic	0.0516
earth	0.0511
painting artstation	0.0507
manga	0.0500
footage	0.0498
photograph	0.0497
fantasy painting	0.0489
intricate details	0.0483
castle	0.0482

Table 3: Top Keyphrases Ranked by Adjusted Scaled Mean SHAP