

Demystifying the Role of Aesthetics in NFT Pricing: Model, Analysis, and Insights

Ahmed Mahrous
King Abdullah University of
Science and Technology (KAUST)
Thuwal, Saudi Arabia
ahmed.mahrous@kaust.edu.sa

Roberto Di Pietro
King Abdullah University of
Science and Technology (KAUST)
Thuwal, Saudi Arabia
roberto.dipietro@kaust.edu.sa

Abstract—The extreme and often seemingly irrational pricing of non-fungible tokens (NFTs) has inspired interest in identifying their valuation determinants. While visual features have been proposed as partial determinants for NFT price variation, previous studies often rely on non-interpretable models and fail to control for a critical confounder: NFT collection identity. In this paper, we analyze over 160,000 NFTs across 32 major collections to evaluate the explanatory power of visual features for NFT pricing. We show that visual models achieve high predictive performance only when collection label leakage is present, falsely inflating results by learning collection-specific price tiers. Once collection identity is controlled, we find that visual features—interpretable or deep-learning based—offer limited generalizability and weak predictive power across collections. However, for a subset of collections, visual features do exhibit modest within-collection predictability, a phenomenon we term *photorelevance*. We introduce a framework to quantify photorelevance and identify two collection-level characteristics—image contrast variance and price variance—as statistically significant predictors of it. Our findings highlight the methodological importance of addressing collection label leakage and suggest that aesthetic attributes play a limited and only localized role in NFT valuation. Code and data are publicly released to enable reproducibility and further research.

Index Terms—Non-Fungible Tokens, NFT pricing, Digital Assets, Deep Learning, Machine Learning

I. INTRODUCTION

In March 2021, a digital artwork NFT by the artist Beeple sold for an unprecedented \$69 million at a Christie’s auction—a sale that stunned the art world and signaled the rapid ascendance of non-fungible tokens into the mainstream [1]. In the same year, the fledgling NFT market ballooned into a multibillion-dollar industry, with overall trading volumes exceeding \$40 billion by the end of 2021 [2]. NFTs—non-fungible tokens—are essentially unique digital assets certified by blockchain technology, verifying their authenticity and ownership [3]. This technological innovation has unlocked new economic possibilities for digital content: from art and music to in-game items and sports collectibles, NFTs enable creators to monetize digital creations and collectors to trade them in open marketplaces.

NFTs derive their value from a complex mix of factors, and understanding what drives NFT prices has become an important question for scholars, artists, and market participants. In traditional art markets, attributes such as an item’s scarcity, reputation of the creator, and market hype or sentiment, all heavily influence prices [4, 5]—and the same appears true for NFTs [6, 7]. For instance, NFTs that have provably rare traits (e.g., a

one-of-a-kind character skin in a profile-picture collection) or that are created by renowned digital artists often command substantial premiums [8]. Community attention and social media “buzz” have also been noted to play a role in NFT valuations [9], as they do with financial markets in general [10]. Furthermore, the potential for illicit activities, including money laundering or wash-trading [11, 12], and other forms of market manipulation common in digital asset markets (e.g., pump-and-dump schemes [13, 14], to artificially drive prices should not be overlooked. Yet, beyond these economic and social factors, a prominent aspect of image-based NFTs demands further scrutiny: the actual visual content of the NFT. Given the provided context, the following questions arise: are there generalizable visual features that consistently aid in predicting NFT prices? Or are buyers and sellers indifferent to an image’s appearance, impacted only by non-visual factors like brand, community, and rarity?

Existing work in this area, somewhat limited in volume to begin with, faces two key limitations. First, most studies to date either focus on small datasets or single collections of NFs or rely on opaque machine learning models with limited interpretability. In the former case, findings about visual determinants might lack generality—a trait that increases price in one collection might not matter in another one. In the latter case, even if a model correctly predicts prices using images, it provides little insight into addressing the why. Second, a methodological concern threatens the validity of previous results: the issue of “collection label leakage”; NFTs are typically grouped into collections, and which collection an NFT belongs to is itself a strong determinant of price. For instance, a random pixel-art Punk from the famed CryptoPunks collection will almost certainly be worth more than a beautiful digital painting from an unknown one, simply due to the CryptoPunks brand [15]. If a predictive model unintentionally learns primarily to identify the collection from the visual data, it might appear to succeed in explaining prices while merely recognizing the image’s collection’s general price levels. We refer to this problem as “collection label leakage”. Many earlier studies do not adequately control for this factor, leaving open the question of how much visual features truly matter.

This paper seeks to address these limitations and provide a more nuanced understanding of the role of visual features in NFT pricing. Specifically, we aim to answer the following research questions:

RQ1: To what extent does collection-label leakage inflate the reported predictive power of visual features for NFT prices?

RQ2: After controlling for collection-label leakage, do generalizable visual features exist that consistently aid

The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST) Center of Excellence for Generative AI - under award number 5940.

in predicting NFT prices across different collections?
 RQ3: How do interpretable, handcrafted visual features compare against non-interpretable, deep learning-based features in their ability to explain NFT price variations?
 RQ4: Given that the effectiveness of visual features for price prediction varies across collections (a concept we term “photorelevance”), are there discernible characteristics of the collections themselves (e.g., visual diversity, image complexity, price variation) that differentiate highly photorelevant collections from those less amenable to visual-only price prediction?

Contributions Our work makes the following contributions:

a) *Experimental Analysis of Collection Label Leakage*:

We show that failing to control for collection label leakage significantly inflates visual-feature model performance by conflating visual style with collection price tiers (RQ1).

b) *Non-existence of Generalizable Visual Features*:

After accounting for collection label leakage, we show that visual features offer no general predictive value for price variations, being weakly effective only in a limited number of more “photorelevant” collections (RQ2).

c) *Interpretability vs. Deep Learning Embeddings*:

Although deep-learning embeddings seem to excel in pooled analyses by picking up collection-specific style cues, neither they nor handcrafted features deliver consistent, generalizable visual-price signals once collection leakage is controlled. However, interpretable metrics enable us to identify specific characteristics that possibly differentiate photorelevant collections.

d) *Analyzing Photorelevance*: We introduce “photorelevance” and identify intra-collection variance metrics that predict when visual aesthetics matter more for NFT pricing (RQ4).

e) *Robust Methodological Framework*: We provide an original NFT-specific methodological framework that corrects for collection leakage, encouraging more robust future research.

f) *Larger Dataset and Reproducibility*: We analyse 160,000+ NFTs across 32 collections, a dataset larger than previous studies using interpretable metrics. All code and data supporting this work are publicly archived¹.

Roadmap The remainder of this paper is organized as follows. Section II reviews existing literature on NFT price determinants, focusing on studies employing visual features, and discusses the methodological oversight of collection label leakage. Section III details our dataset collection, the process of extracting both handcrafted and deep learning-based visual features, and the experimental methodologies employed to reach the findings of this paper. Section IV presents our experimental results, including descriptive statistics, a comparison of model performance under pooled, same-collection, and unseen-collection conditions, an examination of collection-specific performance variability (photorelevance), and an investigation into the collection characteristics that explain this variability. Section VI discusses implications of our findings, acknowledges the limitations of this research, and proposes directions for future work. Finally, Section VII summarizes the main contributions of the paper and offers concluding remarks.

II. RELATED WORK

Research on NFT price determinants has gained traction, driven by the explosive growth of the NFT market. At the

TABLE I: Overview of NFT Price Prediction Studies Utilizing Visual Features

Study	Dataset Size (NFTs)	No. of Collections	Visual Feature Extraction Methods	Visual-Only Accuracy
<i>Studies Using Explainable Metrics (Handcrafted features)</i>				
Guan <i>et al.</i> (2024)	2,895	1	Image Processing Libs	$R^2 = 0.38$
Chen <i>et al.</i> (2023)	19,954	2	Image Processing Libs	$R^2 = 0.24$
Our Work	160,000	32	Image Processing Libs	$R^2 = 0.39$
<i>Studies Without Explainable Metrics (Deep learning)</i>				
Nadini <i>et al.</i> (2021)	4.7M	4,624	AlexNet	$R^2 = 0.08$
Seyhan <i>et al.</i> (2023)	4.7M	4,624	VGG-19, Capsule Net	$F1 = 0.67$
Zheng <i>et al.</i> (2022)	9,045	6	VGG-16	$R^2 = 0.63$
Costa <i>et al.</i> (2023)	4.7M	4,624	ViT, DINO	$F1 = 0.80$
Our Work	160,000	32	CLIP ViT-B/16	$R^2 = 0.36$

heart of this academic curiosity lies a paradox: what drives the extraordinary prices for what, on the surface, are simply arrangements of digital pixels? Scholars have proposed a wide spectrum of variables to account for these seemingly irrational outcomes, including creator reputation, collection or brand affiliation, social media buzz, rarity, speculation (notably, the “greater fool theory”), market sentiment, celebrity or media endorsement, community and network effects, utility, and—more skeptically—manipulative practices such as wash trading or money laundering.

Although the economics of NFT pricing may appear irrational at first glance, established economic theory could offer some insights. The motivations of NFT buyers often echo those observed in traditional art and luxury goods markets, where price is not solely a function of intrinsic utility. In “Veblen goods” markets, for example, demand paradoxically rises as price increases, as goods become status symbols—a phenomenon well documented in behavioral economics. Alongside status signaling, NFT purchasers are also motivated by speculation, desire to support artists, community participation, exclusive rights, and —maybe—aesthetic enjoyment. These multifaceted incentives challenge simplistic notions of rationality and demand a nuanced, evidence-based approach to NFT price prediction.

An ongoing debate in the literature is the role of visual features—the actual content of NFT images—in determining NFT prices. In other words, do the aesthetic features of an NFT actually influence its market price, and if so, to what extent? Several studies have attempted to quantify this effect. In a pioneering study, Nadini *et al.* [15] used convolutional neural networks (CNNs) for image analysis and found that visual features alone explained about 8% of the variance in NFT prices, with collection identity and transaction history being the dominant factors.

Other studies in NFT price prediction also leverage deep learning and machine learning approaches for automated visual feature extraction. Pala and Sefer [16] employed a CNN-based model incorporating not just visual embeddings but also collection identity and influential trader activity, achieving a robust R^2 of 0.655—interestingly, with visual features emerging as the most important factor in their model. Zheng [17] reported that visual features could account for more than 60% of price variation, though this result is likely inflated by collection label leakage and biased by a limited dataset. Shifting from regression to classification, Seyhan and Sefer [18] predicted whether an NFT’s price would rise or fall, and whether it would be resold in the secondary market, reporting F1-scores exceeding

¹All code and analysis scripts used in this study are available at <https://doi.org/10.5281/zenodo.16415242>. The entire NFT image dataset (35GB) and metadata are available at <https://doi.org/10.5281/zenodo.16414740>.

65% based solely on CNN-derived image embeddings. Yet, as with previous work, their results are potentially compromised by failure to control for collection label leakage—a recurring methodological blind spot in the field. Costa et al. [19] also approach NFT price prediction as a classification task, focusing on predicting the price tier (high, medium, or low) for each NFT. They employ Transformer-based deep learning models to extract both visual and textual embeddings from NFT images and their descriptions. Their predictive models achieve high F1-scores: around 80% when using visual features alone, 72% with textual features alone, and 80% when using both modalities.

While deep learning-based models have made substantial contributions to NFT price prediction—often achieving impressive levels of predictive accuracy and demonstrating the power of automated visual and textual feature extraction—two significant limitations are present in the existing literature: lack of interpretability and control for collection label leakage. First, despite their strong performance, these models typically operate as “black boxes”, offering limited interpretability and making it challenging to understand which specific features drive their predictions. This lack of transparency can be a barrier for stakeholders (e.g., artists) seeking actionable insights. Second, a major methodological concern in NFT price prediction research is collection label leakage. This occurs when models, unintentionally, use information about an NFT’s collection during price prediction. Since collection affiliation is a dominant driver of NFT prices, NFTs from popular collections such as Bored Ape Yacht Club or CryptoPunks tend to command higher prices regardless of visual attributes, which can significantly inflate reported predictive accuracy. For example, in a dataset with two collections (e.g., Collection A, average price \$50,000; Collection B, average price \$500), a model may achieve high performance simply by learning to distinguish between collections through visual features unique to each. Consequently, the model does not capture generalizable visual characteristics associated with value, but rather identifies collection membership. This issue persists even if collection names are not provided, as models can infer collection identity from visual patterns alone. As a result, collection label leakage obscures the true explanatory power of visual features, undermining the results of previous work regarding the value of visual aesthetics in explaining NFT prices.

Recognizing the value of interpretability, Chen et al. [20] and Guan et al. [21] adopted handcrafted visual features inspired by principles of art design (e.g., color harmony, composition, edge statistics, and image quality). Chen et al. [20] analyzed two NFT collections, extracting 18 explicit visual metrics from nearly 20,000 NFT images, and applied linear regression and correlation analysis. Their results showed that handcrafted visual features could explain 15–24% of price variation within collections, with moderate correlations (up to 0.30 for number of hues), but also revealed that the significant features differ between collections, limiting cross-collection generalizability.

Guan et al. [21] extend this line of research by comparing the predictive power of handcrafted features and black-box deep learning models using a collection of 3,000 NFTs. While deep models achieve a higher R^2 (up to 0.68), interpretable features still explain about 38% of price variation. They specifically investigate whether aesthetic novelty and image quality serve as a driver of NFT prices. To quantify image quality, the authors first identify optimal values for color saturation and colorfulness—those most strongly associated

with higher prices—and then compute the Euclidean distance from these optima for each NFT. Aesthetic novelty is measured using Integrated Region Matching (IRM) distances, capturing the visual distinctiveness of each NFT in terms of color, texture, and shape relative to other images in the dataset. Their results reveal a nuanced relationship: NFTs exhibiting greater aesthetic novelty tend to command higher prices, but this effect is conditional on high image quality. In other words, the NFT market rewards not just uniqueness, but the combination of novelty with a well-executed design. Such insights highlight the limitations of black-box models, which do not separate the effects of these factors and underscore the value of using interpretable features.

In summary, the existing literature proposes that NFT prices are driven by a variety of economic, social, and visual factors. Table I lists all the papers—to the extent of our knowledge—that report the accuracy of models predicting NFT prices using visuals only. Note that we report our inflated Pooled experiment results, which suffer from the problem of collection label leakage. This is done to match the methodology of comparable work. However, we show that the high results achieved diminish when controlling for collection label leakage.

While several studies have explored the role of image characteristics and found varying degrees of influence, significant methodological challenges persist. Deep learning approaches are not clearly interpretable. Studies employing interpretable handcrafted features, while offering more interpretability, have been limited by small datasets and the number of collections, limiting the generalizability of their conclusions. Previous work also does not adequately control for collection label leakage, an important confounding factor that possibly inflates the reported predictive accuracy of NFT visual features. This paper aims to address these gaps by using interpretable handcrafted features on a larger dataset of around 160,000 NFTs belonging to 32 NFT collections. It also addresses the problem of collection label leakage by comparing the results of a methodology that overlooks it and those that control for it, highlighting the significant inflation of results caused by this often-overlooked confounding factor.

III. DATA AND METHODOLOGY

We devise the following methodology to address the problems discussed in the literature: explainability, sample size, and collection label leakage. NFT metadata and image data were collected using the Reservoir API². The collection process targeted top collections based on trading volume, with a filter requiring collections to have a minimum NFT count of 5,000. Around 5,000 NFTs (with no missing data) were collected for each of the 32 collections, resulting in a total dataset of 160,000 NFTs. For each NFT image, we collected the USD value of the last sale as of 30th April 2025.

As it was observed that NFT price data contains extreme outliers, as shown in Fig. 1c, measures were taken to address that: prices were transformed using a natural logarithm and further winsorized at the 1st and 99th percentiles.

A set of traditional, handcrafted visual features was extracted from each NFT image, leveraging vision libraries such as OpenCV and Scikit-image. To capture potentially useful details not covered by the handcrafted metrics, several pre-trained deep learning-based models were also used to extract embeddings from each image.

²<https://docs.reservoir.tools/what-is-reservoir>

Handcrafted image metrics were chosen to capture interpretable, theory-grounded aspects of visual aesthetics that have been linked to human judgments of aesthetic quality. Texture features (e.g., contrast, sharpness, Haralick GLCM statistics [22], edge density [23], Local Binary Patterns [24], and Histograms of Oriented Gradients[25] capture local spatial patterns related to texture. Structure features (e.g., number of connected components, whitespace percentage, and Hu’s moment invariants [26]) assess image organization and object segmentation.

Simplicity metrics (e.g., unique-color count, edge-based simplicity, shape simplicity, gradient complexity, spatial complexity, and Shannon entropy) quantify perceptual minimalism.

After extracting the handcrafted and deep learning features, three experimental designs were implemented. In particular, the first experiment pools all NFTs together and trains/tests models on this pool of NFTs. In the second experimental design, models are trained and tested on individual collections at a time. This evaluates how a model can predict prices within the context of a specific collection. By design, this experiment does not suffer from collection label leakage. The third experimental design involves training on entire collections and testing on others. In other words, NFTs belonging to 80% of the 32 collections are used for training, while the remaining NFTs are used for testing. This design evaluates models’ ability to generalize to unseen NFT collections, assessing if there are any fundamental patterns relevant to NFTs in general. To enhance robustness, all designs utilized a 10-fold cross-validation. The averages and standard deviations of R^2 scores across the folds were reported. For the second experimental design (same-collection analysis), the average and standard deviation of per-collection average R^2 squares (each collection has 10 folds) were reported.

For the handcrafted features section, five models were tested using the pooled design to determine the most effective model for understanding NFT visual features. The best-performing model was chosen for the three main experimental designs explained above. The same was done for the deep learning sections. The models tested and their performance are reported in Table IV. Images were resized to 224x224 pixels and transformed according to each model’s standard preprocessing pipeline.

To address RQ4 regarding the factors influencing “photorelevance,” an additional analytical step was performed. First, for each of the 32 NFT collections, a photorelevance score was determined by correlating the predicted prices from the same-collection handcrafted features model (as detailed in Section IV-D) with the actual NFT prices within that collection, using Pearson’s correlation r (subsequently Fisher’s z —transformed for more robust regression analysis). Then, for each collection, centrality measures (mean and median) and variance were calculated for each of the individual handcrafted visual features, as well as for the price. This resulted in a dataset where each row represented one of the 32 collections, and columns represented these aggregated collection-level characteristics. These aggregated features were then tested for significant correlation with the collection-specific photorelevance scores. Metrics exhibiting a significant correlation with photorelevance (an absolute Spearman and Pearson correlation greater than 0.3 and a p-value less than 0.1) were selected as initial candidates. To identify the most robust predictors among these candidates, a bootstrap procedure involving 1,000 iterations of Elastic Net regression was employed. Features consistently selected as significant across these iterations were identified, as presented in Table V.

Finally, these selected collection-level features were used in an OLS regression model to predict photorelevance, and a Random Forest model was also trained to assess their importance in a non-linear context (results in tables VI and V, respectively).

IV. FINDINGS

This section presents the experimental results of our study. We begin by illustrating differences in price distributions and visual characteristics between distinct NFT collections, which highlights the context for our subsequent analyses. We then detail the experimental outcomes concerning the predictive power of visual features on NFT prices. By implementing three experimental designs, we demonstrate the issue of collection label leakage and its inflationary effect on pooled model performance. Further results explore model performance variations across individual collections and compare different model architectures. Finally, we address the concept of ‘photorelevance’ and investigate factors influencing it.

A. Preliminary Statistics

We begin by illustrating the heterogeneity inherent in NFT collections with respect to their visual characteristics and price distributions. This helps in understanding the problem of collection label leakage, which is central to subsequent analyses. Fig 1a and Fig. 1b show two NFTs from different collections. Two extracted handcrafted visual metrics, colorfulness and complexity, quantify differences in styles between the collections. The NFT in Fig 1a belongs to the CryptoPunk collection, whose prices are visualized using the top boxplot in Fig 1c. The bottom boxplot belongs to the collection of the other NFT, found in Fig 1b. The boxplots illustrate the significant differences in prices between the collections. It also illustrates the presence of outliers, which have been addressed by log-transformations and winsorization, as explained in Section III. The dot at the extreme right of the boxplot represents a sale of a single NFT image (CryptoPunk #1563, as shown in Fig. 1a, with a sale price of around 57 million dollars³). Fig. 1 is used to explain the collection label leakage problem: a model that learns to classify the NFTs in the figure to their corresponding collections will predict a higher price for the CryptoPunk NFT. This would give the impression that the model learnt to predict prices based on the visuals of an NFT. In reality, the model only learned to classify NFTs into their respective collection. We refer to this problem, which has been ignored by many previous studies, as the NFT collection label leakage problem.

B. Impact of Cross-label Leakage

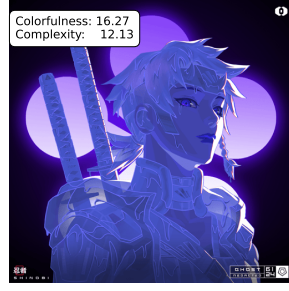
To assess the influence of collection label leakage, a key limitation of previous literature, as explained in Section II and illustrated with Fig. 1, we now present the results of the three experimental conditions defined in Section III: Pooled, Same-Collection (Same-Coll.), and Unseen-Collection (Unseen-Coll.). These conditions allow us to distinguish between apparent predictive power gained from identifying collections versus genuine, generalizable visual feature importance. Table II summarizes the performance (R^2) under each condition.

Table II reveals a dramatic difference in performance across these conditions. In the **Pooled** setting, both the Deep Learning model (CLIP ViT-B/16) and the Handcrafted features model (LGBM), both selected based on Table IV, achieve significant R^2 scores (0.36 ± 0.00 and 0.39 ± 0.01 , respectively). This

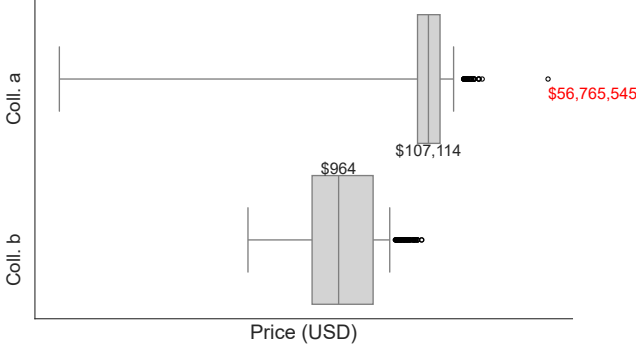
³<https://cryptopunks.app/cryptopunks/details/1563>



(a) An NFT from the CryptoPunk collection.



(b) An NFT from the Ghost Division collection.



(c) Price distributions for the: CryptoPunk (a); and, Ghost Division (b) collections.

Fig. 1: Examples of NFTs and their collection price distributions: (a) an NFT from the CryptoPunk collection; and (b) an NFT from the Ghost Division collection. In (c): Logarithmic-scale boxplots of sales prices across the two collections (CryptoPunk and Ghost Division). Black labels represent median USD prices. The red label is the price of an extreme outlier from the CryptoPunk collection. Values above the 99th percentile are shown as black circles.

suggests that visual features, when considered across the entire dataset without controlling for collection, appear to explain a substantial portion of price variance.

TABLE II: R^2 (Mean $\pm$ Std. Dev.) under pooled, same-collection, and unseen-collection splits.

Model	Pooled	Same-Coll.	Unseen-Coll.
Deep Learning (CLIP ViT-B/16)	0.36 ± 0.00	0.02 ± 0.07	-0.15 ± 0.20
Handcrafted (LGBM)	0.39 ± 0.01	-0.03 ± 0.08	-0.34 ± 0.23

However, this apparent predictive power largely diminishes when controlling for collection labels. In the **Same-Collection** setting, the average performance for the Deep Learning (CLIP ViT-B/16) model drops to an R^2 of 0.02 ± 0.07 , while the Handcrafted features (LGBM) model shows an average R^2 of -0.03 ± 0.08 (an R^2 score below zero indicates the model performs worse than simply predicting the average price within that collection). The standard deviations (± 0.07 for Deep Learning and ± 0.08 for Handcrafted) reflect the variability in performance across the different collections. While the average performance is low, this variability suggests that for some specific collections, visual features might still hold some predictive power, a point explored further in Section IV-C. Overall, these results indicate that features learned in the

pooled setting may have been primarily collection identifiers.

The **Unseen-Collection** results further underscore the lack of generalizable visual predictors across different collection contexts. Both deep learning and handcrafted models perform poorly when tasked with predicting prices for collections entirely excluded from their training data. The deep learning (CLIP ViT-B/16) model yields an R^2 of $(-0.15) \pm 0.20$, and the Handcrafted features (LGBM) model an R^2 of $(-0.34) \pm 0.23$. This inability to generalize, combined with the highly variable and often low performance seen in the Same-Collection setting, strongly suggests that the visual characteristics driving price within one NFT collection do not reliably translate to others.

C. Collection-Specific Performance

The significant standard deviations observed in the Same-Collection results (Table II), particularly for the Handcrafted features model (LGBM), hinted at substantial variability in how well visual features explain prices across different NFT collections. To explore this further, Fig. 2 directly visualizes the R^2 performance achieved by both the Deep Learning and Handcrafted feature models within each of the 32 individual collections. Error bars indicate the standard deviation across the 10 cross-validation folds for each collection-model pair. Moreover, Fig. 2 illustrates the previously suggested heterogeneity. Performance varies, with some collections showing modestly positive R^2 scores (indicating visual features capture some price variance) while others exhibit near-zero or even negative R^2 scores (a negative R^2 indicating the models perform worse than a simple mean price prediction). The small standard deviations across the cross-validation folds for most collections (indicated by the error bars in Fig. 2 and the $\pm$ values in Table III) suggest that these performance levels are relatively stable and not significantly sensitive to specific training samples.

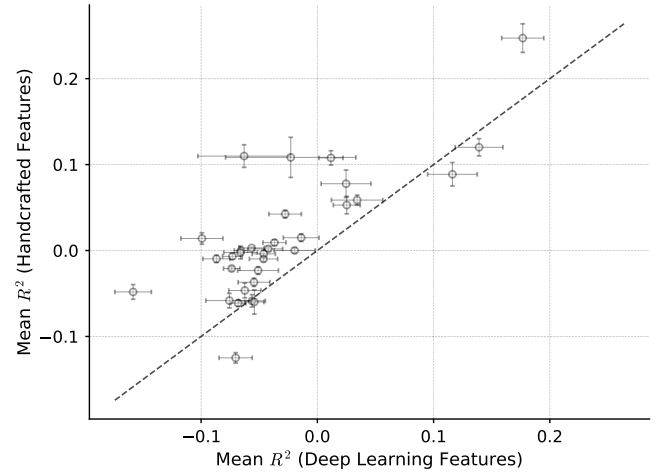


Fig. 2: Per-collection performance of deep learning versus handcrafted features, with error bars showing ± 1 standard deviation across 10 folds.

The inconsistent performance of visual features in explaining price variations across different NFT collections, as illustrated in Fig. 2 and Table III, points to the absence of generalizable visual predictors.

D. Model Performance Comparison (Pooled)

Before proceeding with the main analysis, controlling for collection effects, we evaluated several different model

TABLE III: Top 5 collections by Handcrafted feature performance (R^2 Mean $\pm$ Std. Dev.) under same-collection experiments. Deep learning results are also shown.

NFT Collection	Deep Learning (CLIP ViT-B/16)	Handcrafted (LGBM)
<i>Top 5 Collections</i>		
Azuki_Elementals	0.247 ± 0.033	0.177 ± 0.036
VeeFriends	0.120 ± 0.020	0.139 ± 0.041
Opepen_Edition	0.089 ± 0.027	0.116 ± 0.043
mfers	0.059 ± 0.012	0.034 ± 0.044
BEANZ_Official	0.053 ± 0.020	0.025 ± 0.023

architectures and algorithms on the combined dataset (Pooled design) to select representative models for both the deep learning and handcrafted feature approaches. The results of this comparison are presented in Table IV.

Among the tested deep learning models, which included various architectures—ConvNeXt, ResNet, BEiT, CLIP ViT, and Swin Transformer—the performance was relatively consistent. All models achieved an average R^2 score generally between 0.34 and 0.36. CLIP ViT-B/16, with an R^2 score of 0.361 ± 0.004 , was selected as the representative deep learning model for subsequent experiments.

TABLE IV: Pooled Cross-Validation Performance (R^2 Mean $\pm$ Std. Dev.) for Different Models

Model	R^2
<i>Deep Learning Models</i>	
CLIP ViT-B/16	0.361 ± 0.004
Swin-Tiny	0.351 ± 0.005
BEiT-Base	0.349 ± 0.006
ResNet50 (SimCLR)	0.346 ± 0.003
ConvNeXt-Base	0.345 ± 0.009
<i>Handcrafted Feature Models</i>	
LightGBM (LGBM)	0.386 ± 0.005
XGBoost	0.384 ± 0.006
K-Nearest Neighbors (KNN)	0.243 ± 0.006
Linear Regression (OLS)	0.147 ± 0.006
Ridge Regression	0.146 ± 0.006

For the models trained on the handcrafted visual features, a wider range of performance was observed. Simple Linear Regression (OLS) and Ridge Regression performed poorly, explaining only about $14.7\% \pm 0.006$ and $14.6\% \pm 0.006$ of the variance, respectively. K-Nearest Neighbors (KNN) performed better ($R^2 = 0.243 \pm 0.006$). However, the tree-based ensemble methods, LightGBM (LGBM) and XGBoost, demonstrated significantly stronger performance. LightGBM (LGBM) achieved the highest R^2 score among handcrafted feature models at 0.386 ± 0.005 and was therefore chosen as the representative model for the handcrafted features in the main analyses, whose results are presented in Table II.

It is interesting to note that in this pooled setting, which is susceptible to collection label leakage, the best model using handcrafted features (LGBM, $R^2 = 0.386 \pm 0.005$) slightly outperformed the selected deep learning model (CLIP ViT-B/16, $R^2 = 0.361 \pm 0.004$).

V. EXPLAINING COLLECTION PHOTORELEVANCE

While our previous analyses established that the predictive power of visual features for NFT prices is generally modest and variable across collections once collection label leakage is controlled for, we investigate factors that could explain the variability in “photorelevance” across collections (RQ4). We

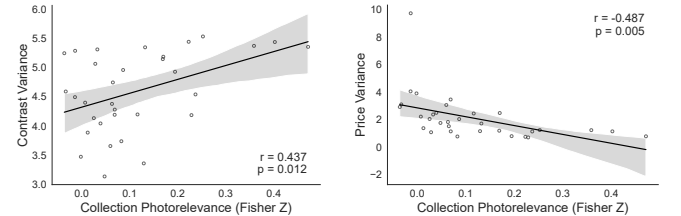
define a collection’s photorelevance as the degree to which its intra-collection NFT prices can be predicted by visual features alone, quantified in our study by the correlation between predicted and actual values obtained from the same-collection handcrafted features model for that specific collection.

While all collection-level metrics (derived from aggregating the handcrafted visual features and price data for each collection, as detailed in Section III) were initially considered, two characteristics consistently emerged as significant across a variety of tests: 1. the variance of visual contrast between NFTs within a collection; and, 2. the variance of prices of NFTs within that collection.

TABLE V: Feature importance comparison. EN sel.: times selected by Elastic Net (of 1000 bootstraps); RF imp.: mean Random Forest permutation importance; r : Pearson correlation with collection photorelevance; p : p-value for the Pearson correlation test.

Feature	EN sel.	RF imp.	r	p
Contrast Variance	857	0.419	0.437	0.012
Price Variance	985	0.305	-0.487	0.005

Table V presents an initial assessment of the importance of these two collection-level characteristics in explaining photorelevance. In a bootstrap procedure involving 1,000 iterations with Elastic Net, both Contrast Variance and Price variance were selected consistently (857 and 985 times, respectively). Furthermore, their mean Random Forest permutation importances were substantial (0.419 for Contrast Variance and 0.305 for Price Variance), suggesting they contribute meaningfully to a non-linear model predicting photorelevance.



(a) Predictability (Fisher’s z) vs. contrast variance ($r = 0.437$, $p = 0.012$). (b) Predictability (Fisher’s z) vs. log price variance ($r = -0.487$, $p = 0.005$).

Fig. 3: Simple OLS fits showing how (a) image contrast heterogeneity and (b) price volatility relate to collection-level predictability.

The Pearson correlation analysis reveals statistically significant relationships. Intra-collection Contrast Variance shows a positive correlation with photorelevance ($r = 0.437$, $p = 0.012$). This relationship, visualized in Fig. 3a, suggests that collections exhibiting greater internal diversity in visual contrast tend to be somewhat more photorelevant. It is plausible that in such collections, a wider spectrum of visual appearances provides more distinct signals that pricing models can leverage.

Conversely, intra-collection Price Variance displays a significant negative correlation with photorelevance ($r = -0.487$, $p = 0.005$), as depicted in Fig. 3b. This indicates that collections with higher price volatility or greater dispersion in prices are less likely to have those prices predictable by their visual features alone. In prices with price outliers or high price variance, it is more probable that non-visual

factors—such as speculative dynamics, rarity, potential illicit behavior, or external market hype—determined the price of some NFTs, overshadowing the role of visual features.

TABLE VI: OLS regression of collection photorelevance on Contrast Variance and Price Variance.

Predictor	Coefficient	<i>p</i> -value
Contrast Variance	0.0809	0.004
Price Variance	−0.0371	0.002

Note: Overall model fit: $R^2 = 0.431$. The overall model significance is tested by the F -statistic ($F = 10.98$), which has an associated p -value ($p_F = 0.000282$).

To evaluate the combined effect of these two characteristics, an Ordinary Least Squares (OLS) regression model was employed to predict collection photorelevance from Contrast Variance and Price Variance. The results of this regression are detailed in Table VI. The model confirms the individual relationships observed: Contrast Variance has a statistically significant positive coefficient (Coef = 0.0809, $p = 0.004$). Price Variance maintains its significant negative coefficient (Coef = −0.0371, $p = 0.002$), indicating that higher price dispersion continues to be associated with lower photorelevance, even when accounting for contrast variance, and vice versa. The overall OLS model demonstrates a statistically significant fit ($R^2 = 0.431$, $F = 10.98$, $p_F = 0.000282$). This implies that approximately 43.1% of the observed variance in photorelevance across the 32 collections can be explained by these two collection-level characteristics.

VI. DISCUSSION, LIMITATIONS, AND FUTURE WORK

Our research provides insights into the role of visual features in NFT pricing. The main finding is that when the collection label leakage is not controlled, the predictive power of visual features for NFT prices is substantially overestimated. This directly addresses RQ1, confirming that much of the apparent price explanation in previous work stems from models learning collection-specific price tiers through their distinct visual styles, rather than identifying generalizable visuals-to-price connections.

When collection-specific effects are isolated (Same-Collection or Unseen-Collection settings), the ability of visual features to predict NFT prices fades. This finding for RQ2 suggests that generalizable visual features that consistently drive NFT prices across different collections are largely non-existent, or at best, have a weak influence. This contrasts with some prior studies reporting high predictive power for visual features (e.g., [17, 19]), suggesting their results may have been inflated by the leakage phenomenon we highlight.

Regarding RQ3, the comparison between interpretable handcrafted features and deep learning embeddings yielded unexpected results (deep learning was expected to achieve superior performance). In pooled scenarios susceptible to leakage, both approaches showed considerable and mostly equivalent predictive power (Table IV). Once leakage was mitigated, neither approach consistently maintained predictive value. The utility of handcrafted features, however, became particularly apparent in analyzing “photorelevance” (RQ4), where their interpretability allowed us to identify specific collection-level characteristics—namely, higher intra-collection visual contrast variance and lower intra-collection price variance—that correlate with the visual features’ ability to

explain prices within a given collection. We find that collections with more visually similar images and higher price variations tend to be challenging for visual-only models to predict. Prices in such collections are more likely to be driven by non-visual factors such as speculation, hype, or wash-trading.

For NFT researchers, our study underscores the methodological necessity of controlling for collection label leakage to avoid spurious conclusions about the importance of visual features. For artists, collectors, and market platforms, it suggests that while aesthetics are part of an NFT’s profile, factors like collection branding, community engagement, perceived rarity, market hype, and—skeptically—market manipulation/money-laundering most likely play a more dominant role in price determination. Visual qualities might be more critical for differentiating value within specific collections where visuals are more appreciated and priced accordingly (i.e., photorelevant collections).

While this study offers a more robust analysis of visual features in NFT pricing, some limitations should be acknowledged. First, our dataset, though large with over 160,000 NFTs from 32 popular (high volume) collections, represents only a fraction of the diverse NFT ecosystem. The findings’ generalizability to collections with different characteristics (e.g., less popular, art-focused collections) or to other types of media (e.g., video, music) is unexplored. Secondly, while our handcrafted visual features were chosen to cover various aesthetic properties, they are not exhaustive. Other, perhaps more domain-specific metrics, could yield different insights. Similarly, while CLIP ViT-B/16 is a powerful pre-trained model, custom-trained deep learning architectures specific to NFT art might perform better, though the core issue of leakage would persist.

The findings and limitations of this study open several avenues for future research. A primary direction is to expand the dataset to include a greater number and diversity of collections. Studying different types of NFTs beyond static images has largely been unexplored. The concept of “photorelevance” also warrants deeper investigation. For instance, it could be interesting to study whether lesser-known collections (which might be less impacted by hype) or collections that are art-driven are more photorelevant. Investigating the specific handcrafted visual features that are significant in predicting prices within photorelevant collections could also offer more granular insights for artists.

VII. CONCLUSION

This study demonstrates the critical impact of collection label leakage, which substantially inflates the perceived predictive power of visual features for NFT prices. Once controlled, we find that generalizable visual features offer non-existent or limited explanatory power across diverse collections, challenging previous work. However, we also introduce the concept of “photorelevance”, identifying that in specific contexts-collections with higher internal visual variance and lower price variance-visual aesthetics can play a more significant role, yet probably still collection-specific, in price differentiation. Our findings advocate for more rigorous, leakage-aware methodologies in NFT research and increase our understanding of NFT valuation, where non-visual characteristics such as brand, community, hype, and—possibly—market manipulation dominate. Finally, further research directions have also been discussed.

REFERENCES

- [1] S. Reyburn, "JPG File Sells for \$69 Million, as 'NFT Mania' Gathers Pace," *The New York Times*, mar 2021, accessed on 25 July 2025. [Online]. Available: <https://www.nytimes.com/2021/03/11/arts/design/nft-auction-christies-beeple.html>
- [2] Chainalysis, "The 2022 Crypto Crime Report," Chainalysis, Tech. Rep., 2022, accessed on 25 July 2025. [Online]. Available: <https://go.chainalysis.com/2022-Crypto-Crime-Report.html>
- [3] Q. Razi, A. Devrani, H. Abhyankar, G. S. S. Chalapathi, V. Hassija, and M. Guizani, "Non-fungible tokens (nfts)—survey of current applications, evolution, and future directions," *IEEE Open Journal of the Communications Society*, vol. 5, pp. 2765–2791, 2024.
- [4] D. Sagot-Duvaroux, S. Pflieger, and B. Rouget, "Factors affecting price on the contemporary art market," in *Cultural Economics*, R. Towse and A. Khakee, Eds. Springer, 1992, pp. 91–102.
- [5] U. Garay, "Determinants of art prices and performance by movements: Long-run evidence from an emerging market," *Journal of Business Research*, vol. 127, pp. 413–426, 2021.
- [6] R. Kräussl and A. Tugnetti, "Non-fungible tokens (NFTs): A review of pricing determinants, applications and opportunities," *Journal of Economic Surveys*, vol. 38, no. 2, pp. 555–574, 2024.
- [7] I. Alon, V. P. Bretas, and V. Katrih, "Predictors of NFT prices: An automated machine learning approach," *Journal of Global Information Management (JGIM)*, vol. 31, no. 1, pp. 1–18, 2023.
- [8] H. Lee, G.-C. Lee, and H.-Y. Koo, "Exploring the relationship between rarity and price of profile picture NFT: A formal concept analysis on the BAYC NFT collection," *Blockchain: Research and Applications*, vol. 5, no. 2, p. 100191, 2024.
- [9] L. Zhang, Y. Quan, J. Cao, K. Z. Zhou, and X. Tong, "Leveraging Social Media Sentiments and Ethical Signals for NFT Valuation," in *2024 IEEE 24th International Conference on Software Quality, Reliability, and Security Companion (QRS-C)*. IEEE, 2024, pp. 206–215.
- [10] A. Mahrous, J. Schneider, and R. Di Pietro, "The role of emojis in sentiment analysis of financial microblogs," in *2023 Fourth International Conference on Intelligent Data Science Technologies and Applications (IDSTA)*. IEEE, 2023, pp. 76–84.
- [11] A. Mosna and G. Soana, "NFTs and the virtual yet concrete art of money laundering," *Computer Law & Security Review*, vol. 51, p. 105874, 2023.
- [12] V. von Wachter, J. R. Jensen, F. Regner, and O. Ross, "NFT wash trading: Quantifying suspicious behaviour in NFT markets," in *International Conference on Financial Cryptography and Data Security*. Springer, 2022, pp. 299–311.
- [13] L. Charfeddine and A. Mahrous, "What drives cryptocurrency pump and dump schemes: Coin versus market factors?" *Finance Research Letters*, vol. 67, p. 105861, 2024.
- [14] A. S. Bello, J. Schneider, and R. Di Pietro, "LLD: A low latency detection solution to thwart cryptocurrency pump & dumps," in *2023 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, 2023, pp. 1–9.
- [15] M. Nadini, L. Alessandretti, F. Di Giacinto, M. Martino, L. M. Aiello, and A. Baronchelli, "Mapping the NFT revolution: market trends, trade networks, and visual features," *Scientific Reports*, vol. 11, no. 1, p. 20902, 2021.
- [16] M. Pala and E. Sefer, "NFT price and sales characteristics prediction by transfer learning of visual attributes," *The Journal of Finance and Data Science*, vol. 10, p. 100148, 2024.
- [17] X. Zheng, "Multimodal learning for improved NFT price prediction," in *2022 IEEE International Conference on e-Business Engineering (ICEBE)*. IEEE, 2022, pp. 74–79.
- [18] B. Seyhan and E. Sefer, "NFT primary sale price and secondary sale prediction via deep learning," in *Proceedings of the Fourth ACM International Conference on AI in Finance*, 2023, pp. 116–123.
- [19] D. Costa, L. La Cava, and A. Tagarelli, "Show me your NFT and I tell you how it will perform: Multimodal representation learning for NFT selling price prediction," in *Proceedings of the ACM Web Conference 2023*, 2023, pp. 1875–1885.
- [20] Y. Chen, Y. Ye, and W. Zeng, "The rich, the poor, and the ugly: An aesthetic-perspective assessment of NFT values," in *Proceedings of the 16th International Symposium on Visual Information Communication and Interaction*, 2023, pp. 1–8.
- [21] C. Guan, D. Ding, J. Ren, and J. Guo, "Unveiling the aesthetic 'wow factor': The role of aesthetic incongruity and image quality in NFT art valuation with computer vision," *Electronic Markets*, vol. 34, no. 1, p. 38, 2024.
- [22] T. Löfstedt, P. Brynolfsson, T. Askund, T. Nyholm, and A. Garpebring, "Gray-level invariant Haralick texture features," *PloS One*, vol. 14, no. 2, p. e0212110, 2019.
- [23] S. Xie and Z. Tu, "Holistically-Nested Edge Detection," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1395–1403.
- [24] M. Pietikäinen and G. Zhao, "Two decades of local binary patterns: A survey," in *Advances in Independent Component Analysis and Learning Machines*. Elsevier, 2015, pp. 175–210.
- [25] L. Zhang, W. Zhou, J. Li, J. Li, and X. Lou, "Histogram of oriented gradients feature extraction without normalization," in *2020 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*. IEEE, 2020, pp. 252–255.
- [26] M. W. Nasrudin, N. S. Yaakob, N. A. A. Rahim, M. Z. Z. Ahmad, N. Ramli, and M. S. A. Rashid, "Moment invariants technique for image analysis and its applications: A review," in *Journal of Physics: Conference Series*, vol. 1962, no. 1. IOP Publishing, 2021, p. 012028.