

Pixels, Scarcity, or Brand? What Visual Features Are Worth in NFT Markets

Samiha Tariq

School of Analytics, Finance, and Economics

Southern Illinois University Carbondale, Carbondale, IL, USA

Abstract

This paper estimates implicit prices for 196 computer-vision descriptors measured on 23,997 images, matched to 93,701 sales of the tokens that carry them across 26 Ethereum collections. What the buyer of a non-fungible token acquires is a ledger record of ownership rather than the image itself, which makes these markets a direct setting in which to ask how much of what people pay for art is payment for how it looks. Studies of art and NFT markets measure images to *predict* prices; none reports what a machine-measured visual feature is worth. The design takes seriously three properties of generative collections: visual style is largely a collection-level attribute, the effective unit of inference is the collection, and scarcity is a deterministic function of the same attribute layers that generate the image. Identifying within collection, controlling for scarcity and conducting inference at the collection level, none of the 196 descriptors carries an identified premium under any standard multiplicity correction. Scarcity does, at 0.154 log points per within-collection standard deviation, and is the only regressor surviving inference at that level. Collection premia are large and move with the market cycle. The pixels are not uninformative: they predict a token's scarcity out of sample, and a visual index built to do so is priced at 0.099 log points on its own and 0.059 against the measured scarcity score, but falls to 18 percent of that and reverses sign once the full attribute vector is included.

JEL: Z11, G12, C23, C58, L86

Keywords: hedonic pricing; non-fungible tokens; computer vision; cluster-robust inference; multiple testing

1 Introduction

A digital artwork is a file. Anyone can copy it, view it, or use it as an avatar without owning it, and the image is identical whether or not they do. What a buyer acquires when they buy a non-fungible token is not the picture but a record on a public ledger saying that this

particular copy is theirs. That makes the market a useful place to ask an old question in cultural economics with unusual precision: how much of what people pay for a work of art is payment for how it looks?

The question is old because hedonic analysis has been asking it of paintings for three decades, and answering it in the only way the data allowed. In an auction of unique works, the attributes are recorded by an expert (genre, medium, period, signature) and the analyst prices those labels. A second strand replaced the expert with a measurement, quantifying colour, palette composition and complexity directly from the image. Both strands face the same difficulty: each painting is a single object that sold once or twice, its attributes are correlated with its maker, and the maker is correlated with everything.

Generative token collections appear to remove that difficulty. Every pixel is recorded rather than described. Provenance is complete and public. Attributes are enumerable rather than interpreted, and the same image sells repeatedly, sometimes dozens of times. If there were ever a setting in which to identify what appearance is worth, this looks like it.

This paper argues that the setting has a structural feature which has gone largely unremarked, and that the feature changes what a token-level hedonic design can identify. **A token in a generative collection is not an independent object.** It is a draw from a layered process in which a small number of attribute categories (a background, a body, a hat, a pair of eyes) are combined algorithmically to produce items that are visually distinct but stylistically coherent. Three consequences follow, and none of them arises for paintings.

First, visual style is largely a property of the *collection*. Tokens within a collection share a palette, a rendering convention and a subject, so a descriptor measured on a token is in substantial part a measurement of which collection it came from, and a premium estimated from pooled variation conflates the two.

Second, the effective unit of inference is the collection, and there are few of them. Only a few dozen collections have traded at a scale that supports estimation. A design with 93,701 sales can rest on a number of independent units small enough that standard asymptotic inference is unreliable.

Third, scarcity is a deterministic function of the same attribute layers that generate the image. A token is rare because it carries uncommon attributes, and it looks the way it does for exactly the same reason. A visual descriptor and a scarcity measure are therefore two readings of one underlying object rather than two independent regressors, and omitting either one does not leave the other undisturbed.

Taking all three seriously changes the answer. It also makes possible a question the literature has not asked. Because attributes generate both the image and the scarcity score, one can ask how much of a token’s attribute structure is recoverable from its pixels alone, and then ask whether anything of the visual signal survives once that structure is conditioned on. The answer to the second question turns out to be no, and the route to it is the contribution of this paper.

Three questions organise the analysis.

RQ1. Do machine-measured visual features carry identified implicit prices within a collection, and are any such premia stable across market conditions?

This is the hedonic question the art-pricing literature has asked of paintings, posed in a setting where the objects are not independent: tokens within a generative collection share a palette, a rendering convention and a subject, so a premium estimated from pooled variation may be measuring the collection rather than the image. Answering it requires identification within collection rather than across, inference at the level at which the features actually vary, and an explicit control for the scarcity that the same attribute layers determine.

RQ2. Does pricing digital art require black-box representations, or do legible descriptors suffice? Deep representations have been used to predict art and NFT prices with success, but always as inputs to a forecasting exercise; what an individual embedding dimension is worth has not been reported. Posing the question as a pricing question rather than a prediction question makes the answer informative in both directions: if embeddings carry premia that interpretable descriptors do not, the loss of interpretability is a price worth paying, and if they do not, it is not.

RQ3. To the extent that visual features are related to price, through what channel does that relation run, and how far does it extend? Images and scarcity are not independent measurements. A token’s attribute layers generate both the picture and the rarity score, so a visual feature may reach price by revealing what a token is made of rather than by being valued in itself. Separating these requires asking what pixels can recover about a token’s attributes, and then whether anything of the visual signal survives once those attributes are conditioned on.

The three empirical patterns reported below are the answers to these questions, in order.

Hedonic analysis of NFT markets is not new. Neither is the use of machine-measured visual attributes to price art: colour, palette composition and complexity have been quantified and priced for paintings and prints for three decades (Holub et al., 1993; Pownall & Graddy, 2016; Stepanova, 2019; Ma et al., 2022; Ju et al., 2019), and for digital art more recently (Alsultan et al., 2024; Chen et al., 2023; Fridgen et al., 2025). Nor is the use of convolutional representations: Nadini et al. (2021) encode 4.7 million NFT images with a pre-trained network, Pala and Sefer (2024) apply transfer learning to NFT price prediction, Aubry et al. (2023) train a convolutional network on 1.1 million auctioned paintings, and Mei et al. (2025) fuse image and tabular data in a multi-modal architecture. Every ingredient of the present analysis is established.

What is not established is what those measurements are *worth*. Across this literature, visual features enter as *predictors* inside a forecasting exercise; **no study reports an implicit price for a machine-measured visual feature estimated under within-collection identification, and none reports one for an embedding dimension at all**. The distinction matters because prediction and identification answer different questions. A feature

can predict prices because it reveals which collection a token belongs to, or how scarce it is, without carrying any premium of its own. Prior work has not been in a position to separate these, and in the settings where most of it was conducted, namely auctions of unique paintings, the separation is not required.

This paper makes three contributions.

First, it establishes what visual features are worth, and through which channel.

Using attribute metadata collected for 23,997 tokens across all 26 collections, it shows that a visual index constructed to predict scarcity is priced (+0.0998 log points), survives a horse race against the measured scarcity score (+0.0590), survives a better-measured scarcity score and a flexible decile control, and is entirely absorbed by the full attribute vector (-0.0108 , bootstrap $p = 0.279$). The absorption is exact by construction rather than approximate: because a sum-of-inverse-frequency rarity score is an additive function of the attribute indicators, it lies in their span, so conditioning on the attribute vector removes scarcity as algebra and not as a proxy control. The implication is a conditional-independence result: *visual features price these assets to the extent that they reveal which attributes a token carries, and no further*. A companion exercise measures how far that revelation extends, and finds it sharply uneven: backgrounds are read almost perfectly from pixels, while eyes and mouths are barely read at all.

Second, it characterises what a token-level hedonic design can and cannot identify. Three features of generative collections, none of which arises for unique artworks, each break a different part of the standard design: visual style is largely a collection-level property, so pooled variation conflates the feature with the brand; the effective unit of inference is the collection, of which there are few; and scarcity is a deterministic function of the same attribute layers that generate the image. The paper shows that these operate separately rather than as one problem, and supplies an analytical result for the third: with token-constant features and a token random intercept, the estimating equations force that intercept exactly orthogonal to the features, so it can absorb only the component of scarcity uncorrelated with them, while the correlated component loads onto the feature coefficients. In a synthetic design with a true feature effect of exactly zero and a correlated omitted regressor, the estimator returns +3.03. This explains why a token-level specification tends to report visual premia; it is a statement about the design, not about any particular study.

Third, it establishes the inference standard such a design requires, and a diagnostic for when that standard binds. Cluster-robust standard errors alone are insufficient here: at 26 collections the asymptotic sandwich flags between four and seven of thirteen features that a wild cluster bootstrap rejects entirely, and one texture descriptor carries an analytic t of +9.32 alongside a bootstrap p of 0.246. A Kish-style concentration diagnostic explains which descriptors fail and why, one number per regressor: the descriptor above draws effectively all of its identifying variation from three collections. The practical claim is sharper than “cluster your standard errors”, at this number of clusters, clustering is

not enough, and which regressors it fails for is predictable in advance.

Two things are deliberately not claimed. The analysis is not offered as a contribution of scale: 26 collections and 93,701 sales are more than most prior NFT hedonic studies use and fewer than several art-market studies, and the binding constraint here is the number of collections rather than the number of transactions. And the results are conditional-independence statements rather than causal ones: visual features are not randomly assigned, and identifying a causal aesthetic premium would require exogenous variation in appearance holding the attribute vector fixed, which observational data on these markets does not contain.

2 Relevant Literature

This section asks one question of the prior literature: for each determinant of value that can be measured from an image, what does existing evidence establish, and in which direction? Table 1 answers it in full and Table 2 records which study estimates which family of determinants. The discussion here traces how the field reached its current position, and identifies where that position is least secure.

2.1 From expert codes to measured pixels

Hedonic pricing (Rosen, 1974) entered art markets as a response to heterogeneity: where objects are one of a kind and change hands rarely, an index that adjusts for quality does better than a simple average (Chanel et al., 1996). Across more than a million auction records, who made a work, what it was made in, and where it was sold all carry stable implicit prices (Renneboog & Spaenjers, 2013), with repeat-sales indices offering a complementary strategy where the goal is an index rather than attribute prices (Mei & Moses, 2002; see Ashenfelter & Graddy, 2003, for a survey). Throughout this tradition the attributes are *expert-coded*: genre, period, medium, size, signature.

A second strand replaced the expert with a measurement. Holub et al. (1993) formalised light and shadow as inputs to price computation. Pownall and Graddy (2016) constructed RGB measures of colour intensity and lightness for contemporary art auctions. Stepanova (2019) quantified palette composition and colour diversity in paintings. Ma et al. (2022) linked measured colour to emotional response and to auction value, in field data and in laboratory experiments across three cultures. Garay et al. (2022a) priced intensity, luminosity and contrast within a single artist’s body of work. Borisov et al. (2022) assembled complexity, segmentation and colour-theoretic features for abstract painting and estimated their implicit prices in a mixed-effects framework.

Read as a body rather than one study at a time, this literature establishes that colour is priced and disagrees about how. **Colour diversity is the only measure on which the evidence points consistently in one direction:** contrastive, many-coloured works command a premium over monochromatic equivalents (Stepanova, 2019; Borisov et al., 2022).

For every other well-studied measure the directions conflict. Saturation is rewarded in Warhol prints (Pownall & Graddy, 2016) and in Basquiat (Garay et al., 2022a) but penalised in CryptoPunks (Alsultan et al., 2024). Lightness carries a discount in contemporary art auctions (Pownall & Graddy, 2016) and in NFTs (Alsultan et al., 2024), yet luminosity is rewarded in Basquiat (Garay et al., 2022a). Complexity is negative in abstract painting (Borisov et al., 2022), positive in CryptoPunks (Alsultan et al., 2024), and changes sign across specifications *within* a single study of 875,389 art NFTs (Fridgen et al., 2025).

These are not disagreements about magnitude. They are sign reversals in the most-studied determinants. Three features of the literature make them unsurprising: samples are typically confined to a single artist, genre or collection; the measures are constructed differently and are rarely benchmarked against one another; and inference is generally conducted at the level of the transaction rather than at the level at which the visual features actually vary. The present paper takes the last of these as its subject.

2.2 Digital art and the promise of complete measurement

Non-fungible tokens appeared to be the setting in which these questions could be settled. Every pixel is recorded, provenance is complete, transaction histories are public, and the trait vector behind an image can be enumerated in full. Art markets rarely supply any of this, and its absence is precisely what has constrained hedonic work on them (Dowling, 2022a; Kräussl & Tugnetti, 2024).

Two designs have been applied. The first uses deep representations as predictors. Nadini et al. (2021) encoded 4.7 million NFT images with a pre-trained convolutional network, reduced the resulting vectors by principal components, and used the leading components to predict sale prices: visual features alone explain between zero and eight percent of the variance, and raise the predictive power of a past-price baseline from 0.55 to 0.60 for secondary sales. Pala and Sefer (2024) applied transfer learning to predict primary sale price, price direction and secondary-sale participation. Ziemke et al. (2023) find that image and text features explain price variation *within* collections but do not carry to collections the model has not seen, which is the pattern Section 6 recovers. The second design measures interpretable aesthetics directly. Alsultan et al. (2024) constructed four aesthetic measures for CryptoPunks and found that investors pay more for colourfulness and textural complexity and less for saturation and brightness. Fridgen et al. (2025) included nine colour shares, a colour count and a normalised entropy measure as controls in a hedonic model of 875,389 art NFTs. Chen et al. (2023) relate nineteen handcrafted descriptors (among them edge, focus-region and blurriness measures) to the listing prices of two collections, Cool Cats and Doodles, and report a weak relationship: the most strongly correlated descriptor reaches 0.245 in one collection and 0.291 in the other, and the descriptors that matter differ between them.

Two things are true of this literature taken as a whole. First, **deep embeddings have been used exclusively for prediction: no study reports an implicit price for an**

embedding dimension. The question of whether pricing digital art requires black-box representations, or whether legible features suffice, has therefore been posed but never answered. Second, whatever the images contribute, identity and transaction history dominate it. Nadini et al. (2021) find that a collection’s own past median price predicts more than half the variance of future sale prices, and that sale history is the strongest predictor throughout. In the painting market, Mei et al. (2025) reach a parallel conclusion using multi-modal deep learning on repeated-sales auction data: artist identity and prior sales history remain the strongest predictors, and visual embeddings add distinct value mainly for works appearing at auction for the first time, where no history exists.

2.3 What generative collections add: scarcity and brand

Two determinants have no counterpart in the painting literature, and both are first-order in NFT markets.

The first is *trait-based scarcity*. Attribute distributions within a generative collection are heavily skewed, and tokens drawn from the thin end of them sell for more, change hands less often, and show higher median returns when resold (Mekacher et al., 2022). Hubbard and LaFave (2026) construct a rarity score as the sum of inverse attribute frequencies for CryptoPunks and document a clear price premium on rarity, alongside a finding that runs against it, in that lighter-skinned tokens sell for significantly more despite being *more* common. Liu et al. (2025) trace a mechanism, showing that within-collection rarity widens the gap between sellers’ willingness to accept and buyers’ willingness to pay, and that sellers concede less on rare tokens. Kong and Lin (2021) and Schaar and Kampakis (2022), as surveyed by Kräussl and Tugnetti (2024), report the same direction. **Scarcity is the one determinant on which this literature is unanimous.** How it should be measured is less settled: Belousov and Yanovich (2025) benchmark rarity meters against market outcomes and find inverse-frequency scores interpretable but not market-optimal.

The second is *collection-level brand*. Nadini et al. (2021) find the collection’s past median price to be the single strongest predictor of an individual token’s price. Fridgen et al. (2025) report large positive coefficients on floor price and trade count and document a superstar effect concentrating sales among a small number of artists. Cui et al. (2023) find rarity and series-level adoption to be the primary drivers in profile-picture collections, with aesthetic signals secondary in the high-price strata.

2.4 A structural feature of generative collections

A token in a generative collection is not an independent object. It is a draw from a layered generative process, in which a small number of attribute categories are combined algorithmically to produce visually distinct but stylistically coherent items. Three consequences follow, and none of them arises for paintings.

First, visual style is largely a collection-level property. Tokens within a collection share a palette, a rendering convention and a subject; Figure 1 shows six collections side by side. A visual feature measured on a token is therefore in substantial part a measurement of its collection, and a premium estimated from pooled variation conflates the two.

Second, the effective unit of inference is the collection, and there are few of them. Only a few dozen collections have traded at a scale that supports estimation. A design with tens of thousands of sales may therefore rest on a number of independent units small enough that standard asymptotic inference is unreliable (Cameron et al., 2008; MacKinnon et al., 2023a, 2023b).

Third, scarcity is a deterministic function of the same attribute layers that generate the image. A rarity score built as the sum of inverse attribute frequencies is, by construction, an additive function of the attribute indicators. A visual feature and a scarcity measure are then two readings of one underlying object rather than two independent regressors, and omitting one does not leave the other undisturbed.

Table 2 shows that no existing study addresses all three. This is not a criticism of that work: the designs in question were developed for markets in which each object is its own unit, where none of the three conditions holds. The open question is what survives when they are transported to a setting where all three do.

2.5 What this paper does

This paper takes the transported design as its object of study. It measures 196 image descriptors (colour, palette structure, composition, edge geometry, line work, texture, image quality, and deep convolutional and style embeddings) across 26 Ethereum profile-picture collections, and asks what a token-level hedonic specification can and cannot identify once the three conditions above are taken seriously: identification within collection rather than across, inference at the collection level rather than the transaction level, and an explicit scarcity control constructed from attribute metadata.

The contribution is threefold. It supplies the head-to-head comparison of legible descriptors against black-box embeddings that the literature has posed but not answered. It shows analytically why a token-level random-effects design cannot be relied upon to absorb correlated scarcity, and what that implies for estimates obtained from such designs. And it establishes, at the token level and free of the small-cluster constraint, the extent to which visual features carry information about a token’s attribute structure, and the extent to which that, rather than appearance itself, is what reaches price.

Table 1: Determinants of value in art and NFT markets: what has been studied, in which direction, and what this paper measures.

Determinant	Evidence base	Direction in prior work	Key citations	Here
A. Colour and palette				
Dominant hue / colour share	Well studied	Mixed and setting-specific. A one-SD increase in blue raises price 10.63% and red 4.20% in painting auctions; in the largest NFT hedonic all nine colour shares enter at ≈ -0.25 with near-identical coefficients, so hues are not differentiated	Ma et al. (2022); Stepanova (2019); Fridgen et al. (2025)	3
Saturation / colour intensity	Well studied	Mixed, and the split runs along the physical-digital line: positive in prints and paintings, negative in the only NFT study. All prior estimates are whole-image; the focal-region variant is row B1	Pownall & Graddy (2016) +; Garay et al. (2022a) +; Alsultan et al. (2024) –	2
Lightness / brightness / luminosity	Well studied	Mixed, majority negative. Darkness carries a premium in contemporary-art auctions and brightness is negative in NFTs; luminosity is positive in Basquiat works	Pownall & Graddy (2016) –; Garay et al. (2022a) +; Alsultan et al. (2024) –	2
Colour diversity / colourfulness	Well studied	Positive and consistent in art markets – contrastive, many-coloured works carry a premium over monochromatic ones – but a simple colour count is ≈ 0 in the largest NFT hedonic	Stepanova (2019) +; Borisov et al. (2022) +; Alsultan et al. (2024) +; Fridgen et al. (2025) null	2
Within-image colour dispersion (range, SD)	Understudied	Positive where the nearest analogues were tested (contrast in hue, contrast in saturation), never estimated in a digital setting	Borisov et al. (2022)	4
Perceptual colour distance (ΔE)	Not yet studied	No prior estimate. Nearest analogue: Itten-theory contrast constructs, which are not perceptual-distance measures	–	1
B. Composition and geometry				

continued on next page

Table 1, continued

Determinant	Evidence base	Direction in prior work	Key citations	Here
Focal-region attributes	Understudied	Estimated on NFT listing prices in two collections; points of interest are insignificant in abstract paintings. No study reports an identified implicit price on realised transaction prices with the attribute vector held fixed	Chen et al. (2023); Borisov et al. (2022), null	3
Edge salience and spatial extent	Understudied	Estimated on NFT listing prices in two collections, where the descriptors that carry the relationship differ between them. No study reports an identified implicit price on realised transaction prices with the attribute vector held fixed	Chen et al. (2023)	5
Line work (thickness, length, curvature, fractal dimension)	Understudied	Mixed. A variance of lines derived by edge detection is significantly positive in a hedonic model of auctioned paintings, with a decreasing marginal effect; a separately constructed lines variance is insignificant in abstract painting	Ju et al. (2019) +; Borisov et al. (2022) null	6
C. Texture and complexity				
Visual / textural complexity	Well studied	Genuinely mixed – the sign is unstable even within one paper. Negative in abstract paintings, positive in CryptoPunks; normalised Shannon entropy runs from -0.006 (n.s.) to $+0.047^{**}$ across six specifications of the same NFT hedonic	Borisov et al. (2022) –; Alsultan et al. (2024) +; Fridgen et al. (2025) unstable	2
Spatial-frequency and LBP texture descriptors	Not yet studied	No prior estimate at this granularity; prior work uses a single scalar complexity measure	–	15

continued on next page

Table 1, continued

Determinant	Evidence base	Direction in prior work	Key citations	Here
Sharpness / image quality	Understudied	A blurriness measure is among the descriptors related to NFT listing prices in two collections; no estimate on realised transaction prices	Chen et al. (2023)	1
D. Deep embeddings				
CNN / pretrained-network embeddings	Studied, but only as predictors	No direction has ever been reported: no study estimates an implicit price for an embedding dimension. On prediction, AlexNet principal components alone reach $R_{adj}^2 \in [0, 0.08]$, and added to a past-price baseline move secondary-sale R_{adj}^2 from 0.55 to 0.60 – useful, but secondary to sale history	Nadini et al. (2021); Pala & Sefer (2024)	100
Style / Gram-matrix embeddings	Not yet studied	No prior estimate in art or NFT pricing	–	50
E. Scarcity				
Trait-rarity score ($\sum 1/f$)	Well studied	Positive and consistent – the most stable finding in the NFT pricing literature. Rarer tokens price higher, trade less often, and earn higher median resale returns	Mekacher et al. (2022); Hubbard & LaFave (2026); Kong & Lin (2021) and Schaar & Kampakis (2022), both as reported in Kräussl & Tugnetti (2024)	1
Attribute count	Understudied	Positive at the extremes, in one collection: zero-attribute CryptoPunks priced 329.56% above average	Schaar & Kampakis (2022) via Kräussl & Tugnetti (2024)	1
Individual attribute identity	Understudied, and all of it CryptoPunks	Attribute-specific, and includes a documented departure from scarcity pricing: lighter-skinned tokens sell for significantly more despite being more common	Hubbard & LaFave (2026); Kong & Lin (2021) and Schaar & Kampakis (2022), both as reported in Kräussl & Tugnetti (2024)	7,342
F. Collection and brand				

continued on next page

Table 1, continued

Determinant	Evidence base	Direction in prior work	Key citations	Here
Collection-level brand premium	Well studied, but as a control rather than a decomposed object	Positive and large. A collection's past median price alone predicts more than half the variance of future sale prices; floor price $\approx +0.93$ and trade count $\approx +1.07$ in log points	Nadini et al. (2021); Fridgen et al. (2025); Cui et al. (2023)	26
Time variation in the brand premium	Not yet studied	No prior formal test	–	160
G. Market conditions and attention				
Crypto-market co-movement	Well studied	Positive but modest and specification-sensitive: Bitcoin shocks raise NFT sales, volatility transmission is measurable but limited, and the crypto-index return is $+0.086^{***}$ in a hedonic yet unstable once selection is corrected	Ante (2022); Brahmana et al. (2025); Dowling (2022 <i>b</i>); Fridgen et al. (2025); Li et al. (2026)	3
Attention, sentiment and social media	Understudied	Negative in the only hedonic estimate, which its own authors call unexpected and attribute to information over-saturation; a study of NFT valuation against traditional asset classes likewise finds sentiment the weakest of the associations it tests	Fridgen et al. (2025); Li et al. (2026)	1
Market regime: are premia state-dependent?	Studied for market-level quantities, understudied for visual-trait premia	Attribute premia are state-dependent where tested, both in Pop Art around 2008–09 and across bull, bear and neutral NFT phases	Garay et al. (2022 <i>b</i>); Kang & Lee (2025); Özdemir & Kumar (2024); Tang et al. (2023)	13

Notes. **Evidence base** is assigned by a fixed rule: *well studied* = three or more independent studies estimate a price relationship for the determinant, at least one in a digital setting; *understudied* = one or two studies, or evidence only from physical art markets, or the determinant enters only as a control; *not yet studied* = no study in the assembled literature estimates a price relationship for it, in which case the nearest analogue is named. The assembled literature is the 27 empirical studies in Table 2. **Here** counts the regressors, or for fixed effects the units, in this paper's design that measure the determinant; the full feature list is in Appendix Table 17. Rows A1–D2 are a strict partition of the 196 image features, so their counts sum to $46 + 150 = 196$; features that could sit in two rows are assigned to the row prior work actually estimates.

Signs after a citation give the direction that study reports.

Table 2: Which study estimates which determinant. Modelled on the study $\times$ method grid of Kräussl & Tugnetti (2024).

Study	Setting	Col.	Geom.	Text.	Emb.	Rar.	Brand	Mkt.
Pownall & Graddy (2016)	Contemporary art	●	—	—	—	—	○	—
Stepanova (2019)	Paintings	●	—	—	—	—	○	—
Ma et al. (2022)	Paintings, lab	●	—	—	—	—	○	—
Garay et al. (2022 <i>a</i>)	Basquiat	●	—	—	—	—	—	○
Garay et al. (2022 <i>b</i>)	Pop Art	○	—	—	—	—	—	●
Borisov et al. (2022)	Abstract paintings	●	●	●	—	—	○	—
Nadini et al. (2021)	NFTs, 6 categories	—	—	—	●	—	○	○
Pala & Sefer (2024)	NFTs	—	—	—	●	○	—	—
Ziemke et al. (2023)	NFTs, OpenSea	—	—	—	●	○	●	○
Alsultan et al. (2024)	CryptoPunks	●	—	●	—	○	—	○
Chen et al. (2023)	NFTs	●	○	●	—	—	○	—
Fridgen et al. (2025)	875,389 art NFTs	●	—	●	—	○	●	●
Mekacher et al. (2022)	410 collections	—	—	—	—	●	○	—
Hubbard & LaFave (2026)	CryptoPunks	○	—	—	—	●	—	○
Kong & Lin (2021)	CryptoPunks	—	—	—	—	●	—	●
Schaar & Kampakis (2022)	CryptoPunks	—	—	—	—	●	—	—
Cui et al. (2023)	PFP NFTs	○	—	—	—	●	●	—
Kireyev & Lin (2021)	CryptoKitties	—	—	—	—	●	○	—
Goldberg et al. (2021)	Decentraland	—	●	—	—	—	—	—
Dowling (2022 <i>a</i> , 2022 <i>b</i>)	Virtual land, NFTs	—	○	—	—	—	—	●
Ante (2022)	NFT market	—	—	—	—	—	—	●
Tang et al. (2023)	Ethereum NFTs	—	—	—	—	—	—	●
Özdemir & Kumar (2024)	4 NFT markets	—	—	—	—	—	○	●
Brahmana et al. (2025)	5 NFT assets	—	—	—	—	—	—	●
Li et al. (2026)	NFT vs. asset classes	—	—	—	—	—	—	●
Kang & Lee (2025)	NFTs	○	—	—	—	○	○	●
This paper	26 collections	●	●	●	●	●	●	●

Notes. • the study estimates a price relationship for this determinant family; ◦ the family enters only as a control or a descriptive statistic; – absent. Column headings abbreviate colour and palette, composition and geometry, texture and complexity, deep embeddings, rarity and scarcity, collection and brand, and market conditions. Findings for Kong & Lin (2021), Schaar & Kampakis (2022), Kireyev & Lin (2021) and Goldberg et al. (2021) are as surveyed in Kräussl & Tugnetti (2024).

3 Data

3.1 Collections and Transaction Data

The sample covers 26 Ethereum profile-picture collections: generative projects in which layered visual components are combined algorithmically to produce items that are visually distinct but stylistically coherent. The collections were selected from a ranking of the largest collections by cumulative secondary-market volume, and their supply ranges from 5,555 to 21,931 tokens.

Transactions are taken from the public Ethereum record. Because transfers are recorded on chain regardless of where a trade was arranged, the sample covers sales settled on any venue rather than on a single marketplace, and each observation carries its price, its counterparties and its timestamp. The venue of an individual sale is not identified, so marketplace-specific fee and royalty regimes are not controlled for directly; to the extent that venue shares shift over the period, they are absorbed into the month effects.

The study window runs from January 2021 to March 2025. Prices are realised, market-clearing transaction prices, originally denominated in ETH and converted to USD at the trade-day rate.

3.2 Sampling

The collections. The sample is drawn from the largest Ethereum collections by cumulative secondary-market volume as of March 2025, restricted to *generative character collections*: projects in which layered visual components, a background, a body, clothing, headwear and facial features, are combined algorithmically to produce visually distinct but stylistically coherent depictions of a figure. The restriction is what the design requires, since identification comes from visual variation *within* a collection, holding artist, brand and generative process fixed. Collections of other kinds are not eligible on that criterion: virtual land, naming and utility assets, curated single-edition art platforms, abstract generative series with no depicted subject, trading-card and gaming assets, and wrapper contracts.

The resulting sample is 26 established, actively traded collections, and this is the segment of the market in which the paper’s question is well posed: the segment where thousands of tokens share an artist, a brand and a generative process, and differ only in the attributes drawn for them. The claims below are stated for that population. The criterion is close to exhaustive among the very largest collections and becomes selective further down the ranking,

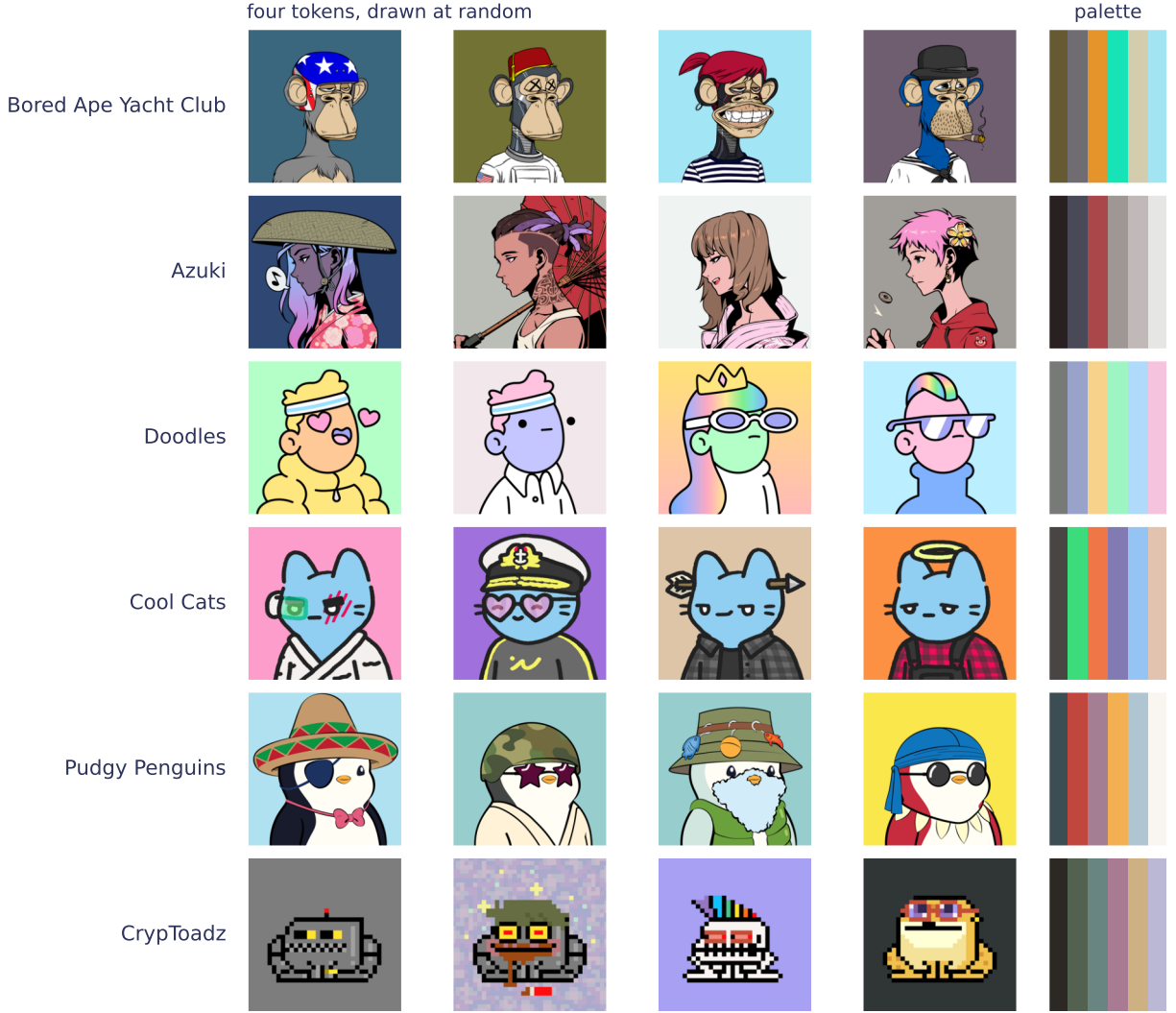


Figure 1: **Six of the twenty-six collections.** Each row shows four tokens drawn at random from one collection, with the colour palette of that collection alongside. Within a row the tokens differ in the attributes drawn for them, in headwear, clothing, expression and background. Across rows the visual style changes completely, and so does the palette. This is the first of the three properties set out in Section 2.4, and it is why every specification in this paper identifies from variation *within* a collection rather than across collections: a descriptor that separates a Doodle from an Ape is measuring the collection, not the token. Palettes are six-colour k -means summaries of forty images per collection, ordered from dark to light so that rows are comparable. Token draws use a fixed seed and are reproducible from the accompanying code.

where the number of collections that could be assembled was bounded by time rather than by any property of the collections themselves. Three further collections were sampled but are not used, because usable images were not available for them.

Nothing in the estimates depends on how the collections were chosen. Selection operates at the collection level and uses no token-level information, while every specification identifies from variation within collections with collection effects absorbed. A collection-level selection rule is therefore orthogonal to the object being estimated by construction: it cannot induce, inflate or mask a relationship between a token’s visual features and its price relative to the other tokens of the same collection.

The tokens. Within each collection a random sample of tokens was drawn without replacement under a fixed seed, and the complete history of every drawn token is observed. No token characteristic plays any part in the draw: not scarcity, attribute composition, appearance, age or prior trading activity. Because the unit sampled is the token rather than the sale, statistics computed over sales are unbiased for the underlying population by construction. Sampling sales instead would over-weight frequently traded tokens and would make any statistic involving turnover a function of the sampling scheme.

Tokens that never trade. Of the tokens sampled, 22.3% carry a mint record and no subsequent sale, and a further 2.8% show no recorded activity; the remaining 74.9% traded at least once and form the pool from which the estimation sample is built. This is a property of the market: a large share of supply in these collections is minted and held, and the share that has never resold ranges from 2.2% to 72.2% across collections. A token that never sold contributes no price observation, in the same way that an unsold house contributes nothing to a housing hedonic, and no sale of a sampled token is discarded. Collections whose tokens turn over rarely therefore contribute proportionally fewer observations, a compositional difference that collection effects absorb.

3.3 Attribute Frequencies and Measurement

Two features of the token sample bear on how scarcity is measured, and both are reported rather than assumed.

The share of a collection that is sampled varies with its size. Collection supply ranges from 5,555 to 21,931 tokens, a factor of 3.9, and the share of each collection observed ranges from 4.3% to 12.9%, with a median of 10.0%. Table 3 reports the share for every collection.

What a token sample does damage is the rare tail of the attribute distribution. An attribute value held by a handful of tokens may not be drawn at all, so frequencies computed within the sample overstate the rarity of common values and understate it for values that were missed. This is measurable, because full-population attribute counts are available for 18 of the 26 collections. Comparing rarity computed on sampled denominators against rarity computed on population denominators gives a rank correlation with a median

of 0.931 across those collections, but a range from 0.985 down to 0.271, and 1,270 attribute values that exist in the population and appear nowhere in the sample. The worst case is a collection in which 739 of 1,036 population attribute values were never drawn.

The direction of the resulting bias is known. Sampled denominators produce classical measurement error in the scarcity regressor, which attenuates its coefficient toward zero. The reported scarcity premium is therefore a *lower bound*, and the null result on visual features is conservative for the same reason: a regressor measured with error absorbs less of the variation it should, leaving more available to the features tested alongside it. Re-estimating on population denominators for the 18 collections where they are available raises the scarcity coefficient from +0.1542 to +0.1630, and excluding the five collections with near-degenerate attribute distributions raises it to +0.1657. Both movements go in the direction attenuation predicts.

Finally, the binding constraint is the number of collections, not the number of tokens. Modelling the full population of all tokens within these same 26 collections, roughly eleven times the current sample, moves no visual feature from insignificance to significance, and the median share of a feature’s standard error that additional tokens could ever remove is 13.6%. The remainder is dispersion of the coefficient across collections, which more tokens within the same collections cannot touch. On this evidence the scale of the transaction sample is not what determines the paper’s conclusions, and enlarging it would not change them.

Table 3: The 26 collections: supply, sampling and rarity measurement.

Collection	Supply	Sampled	Rate	Sales	Rank corr.
lil_pudgys	21,931	1,049	4.8%	3,414	0.864
nakamigos	20,000	858	4.3%	4,826	0.919
beanz	19,950	1,412	7.1%	5,789	0.931
clone_x	19,764	1,327	6.7%	4,491	0.880
mutant_ape	19,567	1,314	6.7%	6,330	0.970
hashmasks	16,384	1,211	7.4%	2,853	0.964
veefriends	10,255	494	4.8%	1,052	0.271
mfers	10,021	1,175	11.7%	4,478	0.981
creatures	10,000	747	7.5%	2,439	0.923
azuki	10,000	1,120	11.2%	5,215	0.910
moonbirds	10,000	902	9.0%	2,806	0.940
bored_ape	10,000	1,135	11.3%	5,805	0.960
doodles	10,000	1,096	11.0%	4,722	0.940
cool_cats	9,999	1,094	10.9%	5,225	0.930
pudgy_penguin	8,888	1,112	12.5%	6,522	0.925
on1_force	7,777	961	12.4%	4,304	0.985
cryptoadz	7,025	861	12.3%	3,391	0.928
karafuru	5,555	714	12.9%	2,422	0.935

continued on next page

Table 3, continued

Collection	Supply	Sampled	Rate	Sales	Rank corr.
<i>Total supply unavailable; sampled denominators retained</i>					
creepz	–	214	–	306	–
degods	–	58	–	208	–
kanpai_pandas	–	594	–	1,341	–
milady	–	654	–	2,000	–
phantabear	–	988	–	2,650	–
renga	–	878	–	3,393	–
sappy_seals	–	871	–	4,042	–
wow	–	1,158	–	3,677	–
Total		23,997		93,701	

Notes. Supply is the collection’s total issued supply, which is unavailable for eight collections; those are reported without it. “Rank corr.” is the Spearman correlation between a token’s rarity computed on sampled attribute frequencies and on full-population frequencies, available only where total supply is known. The share of a collection observed varies because the number of tokens drawn does not scale one for one with supply, and supply ranges by a factor of 3.9. **The lowest rank correlation, 0.271, belongs to a collection in which 739 of 1,036 population attribute values were never drawn;** it is retained on sampled denominators throughout and its rarity values should be read as noisier than the rest. Sales are those in the final estimation sample.

3.4 Prices and Sample Construction

The dependent variable is $\log p_{it}$, the log of the realised sale price.

Two screens are applied. Events without a positive price, that is mints and non-priced transfers, are removed. Sales below 5% of their collection’s trailing seven-day median are removed as unlikely to be arm’s-length: at the low extreme these are transfers between related wallets, private arrangements and wash trades rather than market-clearing prices. The second screen removes 206 sales, 0.22% of the sample. Table 4 reports the construction in full.

Market controls are the daily returns on ETH, BTC and SOL (Yahoo, n.d.), the NASDAQ composite return (Federal Reserve Bank of St. Louis, n.d.), and a cryptocurrency sentiment index (Alternative.me, n.d.). These are the classes that move NFT valuations most in the wider literature: Li et al. (2026), testing NFT prices against stocks and bonds, exchange rates, commodities and cryptocurrencies, find cryptocurrencies and equities the strongest associations and sentiment the weakest. Equity index returns are defined only on trading days while these tokens trade every day, so index levels are carried forward through non-trading days and equity returns set to zero; returns are deliberately not interpolated.

3.5 Image Feature Extraction

Each token’s rendered image is reduced to 196 numerical descriptors in eight families. Forty-six are interpretable measurements of named visual properties: fourteen colour and palette measures (mean hue, saturation and lightness, palette dispersion, colourfulness, a perceptual colour distance), three measures of the colour of the salient region, five measures of edge extent and spatial framing, six measures of line work including thickness, curvature and fractal dimension, fifteen texture descriptors from spatial-frequency bands and local binary patterns, and three image-quality measures of entropy, energy and sharpness. The remaining 150 are deep representations, described in Section 3.6.

No screening step is applied. Every one of the 196 descriptors is carried into the analysis and tested individually. A screening step that retains a compact subset invites the objection that the subset was chosen well or badly, and it makes the resulting null conditional on a selection rule that the reader cannot inspect. Testing all 196 and correcting for multiplicity is more demanding, and the correction is reported four different ways in Section 5.4 so that the result does not depend on which is preferred.

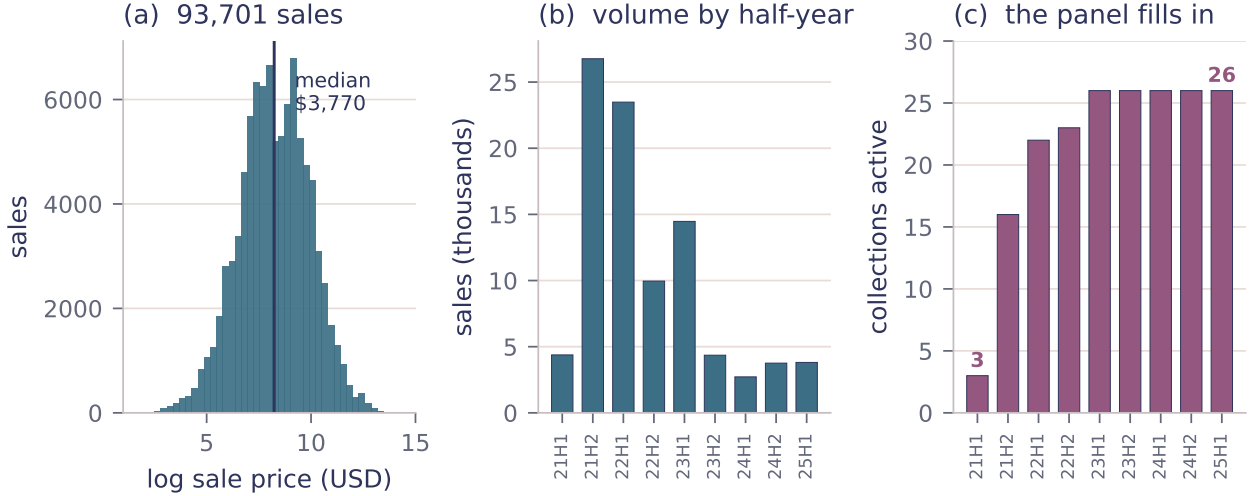


Figure 2: **The estimation sample.** Panel (a) is the distribution of the dependent variable across the 93,701 sales that survive the floor-relative screen. Panel (b) is trading volume by calendar half-year; the 2021 boom and the subsequent contraction are both visible, which is why every specification carries month fixed effects. Panel (c) is the number of collections with at least one sale in each half-year. The panel is unbalanced by construction: three collections are active in 2021H1 and all 26 only from 2023H1. Any statistic computed across collections within an early period is therefore taken over very few units, which is why cross-collection dispersion is not compared across periods anywhere in this paper.

Notes. Equity-index returns are defined only on trading days, while these tokens trade every day of the week. A complete-case rule on those series would therefore remove all 26,871 weekend sales in

Table 4: Sample construction.

Step	Sales	Reason
Recorded sales	94,039	
Less non-priced events	93,950	mints and events with price ≤ 0
Less duplicated feature rows	93,914	two tokens carried four image-feature vectors each
Less missing controls	93,907	Specification H
Less floor-relative screen	93,701	below 5% of the trailing 7-day collection median
Less tokens without rarity	93,383	estimation sample with scarcity

the window and no weekday sales at all. Index levels are instead carried forward through non-trading days and equity returns set to zero, on the ground that no new equity information arrives; returns are deliberately not interpolated. The screen removes 206 sales (0.22%). Deduplication collapses only groups whose image features disagree; deduplicating on identifiers and price alone would remove 6,972 rows (7.4%), because genuine repeat sales at the same day-resolution date and price are not duplicates.

3.6 Deep Visual Embeddings

Forty-six of the 196 descriptors measure a named and local property of the image: the mean of a colour channel, the energy in a spatial-frequency band, the share of detected line pixels that are curved. Each is a statistic of one operator applied to the whole frame, and each is legible in the sense that a coefficient on it can be read as a premium on a design choice an artist could have made differently. What a buyer confronts, however, is not forty-six statistics but a single image, and there is no reason to assume that willingness to pay is a linear function of marginal summaries of it. Two properties that a collector plausibly does respond to are *configural*: they are determined by how the parts of an image are arranged relative to one another and are not recoverable from any individual primitive. The first is *subject and figure composition*: what is depicted, and how it is placed and scaled within the frame. The second is *texture-and-style signature*: the manner of execution, independent of what is depicted. Both have long-standing counterparts in the painting literature, where subject matter and school or manner are priced routinely, and neither has a handcrafted counterpart in the feature set assembled here: a mean hue does not know whether it belongs to a face or to a background, and a fractal dimension does not distinguish a hand-drawn stroke from a vector-rendered one.

Deep representations are included because each is a machine proxy for exactly one of those two properties. A convolutional descriptor, read from a network trained to discriminate objects, summarises what an image contains and how it is arranged: it is a subject-and-composition measurement. A Gram matrix, formed from correlations between convolutional channels within a layer, discards spatial position by construction and retains only the joint distribution of local textures; this is the standard formalisation of artistic style.

The specific representations used are a 2,048-dimensional ResNet-50 global descriptor and a 348,160-dimensional stack of VGG-19 Gram matrices taken at four convolutional layers. **The reason for including them is a specification argument, not a predictive one.** If either configural property enters willingness to pay and is not spanned by the forty-six primitives, then a hedonic containing only the primitives is misspecified, and the premia estimated on the interpretable features absorb part of what the omitted property is worth. Including the embeddings is therefore a test of the interpretable specification rather than an attempt to improve its fit, and, because no study in either the art or the NFT literature has ever reported an implicit price for an embedding dimension, it is a test that has not previously been run.

The raw representations cannot enter a hedonic estimated across 26 collections in their native form: 350,208 columns against a design whose binding constraint is the number of clusters. They are therefore collapsed by principal components, retaining 100 convolutional and 50 style components. Two features of that reduction matter for what can be concluded from it. It is **unsupervised** (computed from images alone, with no reference to price), so it cannot manufacture a pricing relationship, and a null obtained after it is not an artefact of having selected on the outcome. But it is also **lossy in a specific direction**: a principal component is a direction in a representation space, not a nameable property, so a coefficient on one cannot be read as a premium on any design choice. This is the substantive cost of the collapse, and it is why the embeddings are assessed here as a block against the interpretable block, rather than component by component.

Assessed on both margins, the embeddings did not outperform the interpretable features in this benchmark. On implicit prices, the full sweep of all 196 descriptors (each entered individually with scarcity controlled, collection and month effects absorbed, a wild cluster bootstrap at $G = 26$ with $B = 9,999$ replications and a Romano–Wolf stepdown across all 196 candidates on common draws) returns no survivor at the five percent level from either set. The strongest embedding is a style component, at -0.0163 with a bootstrap p of 0.0008 that becomes 0.346 once the search over 196 candidates is accounted for; the strongest convolutional component reaches $+0.0083$, bootstrap $p = 0.0149$, adjusted $p = 0.997$. The null is symmetric: no interpretable feature survives either. On out-of-sample prediction, with folds grouped by token because image features are token-constant and most sales are repeat sales of the same token, the 46 interpretable features reach an R^2 of 0.417 against 0.124 for the 150 embeddings under gradient boosting, and 0.223 against -0.004 under a penalised linear model; collection identity alone reaches 0.425 . **Two qualifications on that comparison should not be dropped.** No family was separately tuned, and the families range from three columns to a hundred, so the ranking records usefulness under one common configuration rather than inherent merit. And the embeddings are the block most damaged by moving from random to token-grouped folds (0.317 falls to 0.124 , the largest such gap of any block tested), so most of what they appeared to contribute under a random

split was memorisation of token price levels rather than visual information.

There is one task at which the embeddings clearly win, and it identifies what they are measuring. Asked to recover *which* attribute value a token carries from its pixels alone, across 138 collection-by-trait-type pairs, the convolutional and style blocks reach median skill scores of +0.811 and +0.797 on the detectable pairs, against +0.294 for texture and +0.262 for colour, with one exception that is instructive, background colour, where the fourteen colour descriptors alone reach +0.909 and beat both embedding blocks. Asked instead *how rare* a token is, the ranking reverses: the interpretable block reaches +0.164 and the embeddings +0.083. The two results are consistent rather than contradictory, and together they say what each set of features is for. **The interpretable descriptors measure how much of something is present in an image; the embeddings identify what that something is.** Rarity, in these collections, is a matter of how busy an image is, which the interpretable measures capture directly; attribute identity is fine-grained object recognition, which is the task convolutional networks were built for. The embeddings are therefore not weak representations. They are strong representations of a property that turns out not to be separately priced.

Finally, the mutual orthogonality of principal components is a property of the block they are extracted from and does not extend across blocks, and no embedding was ever compared against a handcrafted descriptor when the descriptor set was assembled. The relationship between the two sets is accordingly measured rather than assumed. Across all 6,900 cross-family pairs the largest absolute correlation is 0.31 within collection and 0.23 in levels, so a cross-family screen at the same threshold would have removed nothing. Because pairwise non-redundancy does not imply orthogonality, each interpretable feature was also regressed on all 150 embeddings jointly: the median R^2 is 0.056 within collection, and the largest, for perceptual colour distance, is 0.212. The two sets are close to orthogonal without being orthogonal, and the embeddings span too little of any interpretable descriptor to be absorbing variation that belongs to it. That is weaker than orthogonality, and it is what the design actually requires.

3.7 Attribute Metadata and the Scarcity Measure

Scarcity is constructed from the attribute metadata that underlies each image. Every token carries a full attribute vector, listing its background, body, headwear, eyes and so on, and this is available for 23,997 tokens across all 26 collections, or 99.4% of the sampled tokens.

The scarcity score is the sum of inverse attribute frequencies,

$$R_i = \sum_k \frac{1}{f_{c,k}(a_{ik})}, \quad (1)$$

where $f_{c,k}(a_{ik})$ is the share of tokens in collection c carrying value a_{ik} in attribute category

k . The score is then z -scored within collection, so that it measures how unusual a token is relative to its own collection rather than across collections of different sizes and conventions.

Two properties of (1) matter later. The first is that z -scoring within collection makes $E[R_i | c] = 0$ by construction, so no analysis of this measure can earn explanatory power from recognising which collection a token belongs to. The second is that the score is an *additive* function of the attribute indicators: each term is a constant given the attribute value, so R_i lies exactly in the span of a full set of attribute dummies. Conditioning on that full set therefore absorbs scarcity exactly, as algebra, without assuming any functional form. That is what makes the test in Section 6.3 possible.

On the choice of measure. Inverse-frequency scores are not the best available predictor of market outcomes; measures fitted directly to trade data outperform them (Belousov & Yanovich, 2025). That is not the criterion here. The measure is chosen because it is additive in the attribute indicators, and a better-fitting but non-additive alternative would forfeit the absorption argument entirely. The results below are also consistent with that ranking: Section 6.3 shows the scalar score to be a lossy summary of the attribute vector, which is precisely why a measure fitted to market data improves on it.

Measurement. Attribute frequencies are computed from the sampled tokens, so an attribute value held by very few tokens may be under-observed. Section 3.3 quantifies this and shows that the resulting classical measurement error attenuates the scarcity coefficient toward zero, making the reported premium a lower bound.

3.8 Market Controls and Multicollinearity

The specification includes five market controls: the daily returns on ETH, BTC and SOL, the NASDAQ Composite return, and the Fear & Greed index. Because cryptocurrency returns co-move strongly, we report the full correlation matrix and variance inflation factors on the estimation sample ($N = 93,701$) in Table 5. The ETH–BTC correlation is the largest of the ten pairs at $r = 0.83$, and ETH–SOL and BTC–SOL follow at 0.66 and 0.61. This co-movement is real, but it is not severe enough to threaten the estimates: the largest variance inflation factor is 3.66 (ETH), followed by 3.36 (BTC), and every control lies well below the conventional thresholds of 5 and 10. The associated tolerances never fall below 0.27, the largest standard-error inflation is a factor of 1.91, and the condition number of the control correlation matrix is 16.5.

Two points about interpretation. First, multicollinearity inflates standard errors but leaves coefficients unbiased, so it bears on the precision of the individual crypto coefficients, not on the visual-feature estimates that carry the paper’s argument; the image features are near-orthogonal to the market controls, which are common to all tokens traded on a given day and are absorbed in the month fixed effects. Second, the reader may reasonably ask whether the ETH–BTC correlation is why the BTC return is not individually identified in the estimates reported below ($t = -0.4$, wild cluster bootstrap $p = 0.93$). It is not. Collinearity

does inflate the BTC standard error by a factor of 1.65 relative to a specification without ETH and SOL, but assigning BTC that uninflated standard error while holding its conditional point estimate fixed still yields $|t| = 0.74$. The BTC null is driven by a conditional coefficient that is essentially zero (-0.004), not by variance inflation. The BTC coefficient does become significant if ETH and SOL are dropped, but it then moves by a factor of 5.6, because BTC is proxying for the common crypto-market factor those returns otherwise absorb, a change of estimand rather than a recovery of precision. The ETH coefficient, by contrast, is stable and significant whether or not BTC is included (-0.056 versus -0.059), so the ETH–BTC correlation does not compromise identification of the ETH effect.

For completeness, the one pair in the original control set that did breach conventional thresholds was the equity pair. The S&P 500 and NASDAQ returns correlate at $r = 0.9524$ on this sample, giving variance inflation factors of 10.8 and 11.1 respectively when both are included. This is why the reported specification retains the NASDAQ Composite alone; the image coefficients are unchanged to four decimal places by the choice.

Table 5: Correlations and variance inflation factors, market controls

<i>Panel A: pairwise correlations</i>					
	ETH	BTC	SOL	NASDAQ	F&G
ETH return	1.000				
BTC return	0.828	1.000			
SOL return	0.661	0.611	1.000		
NASDAQ return	0.450	0.463	0.395	1.000	
Fear & Greed	0.097	0.068	0.120	0.103	1.000

<i>Panel B: variance inflation</i>			
	VIF	Tolerance	$\sqrt{\text{VIF}}$
ETH return	3.66	0.273	1.91
BTC return	3.36	0.298	1.83
SOL return	1.85	0.540	1.36
NASDAQ return	1.32	0.757	1.15
Fear & Greed	1.02	0.980	1.01

Notes. Estimation sample, $N = 93,701$ sales. VIF for control j is $1/(1 - R_j^2)$ from an auxiliary OLS of j on the remaining four controls and a constant; tolerance is $1 - R_j^2$. Correlations and VIFs are invariant to the scaling of the individual regressors. No control exceeds the conventional thresholds of 5 or 10. The condition number of the Panel A matrix is 16.5. The S&P 500 return is excluded from the specification: it correlates with the NASDAQ return at 0.9524, and including both raises their VIFs to 10.8 and 11.1.

4 Empirical Strategy

Section 2.4 identified three properties of generative collections that a token-level hedonic design must confront. This section states the specification, and then addresses each property in turn: the level at which the premium is identified, the level at which inference is conducted, and what a token random intercept can be relied upon to absorb.

4.1 The hedonic specification

Let p_{it} be the realised price of token i in month t , x_i the vector of image descriptors measured from the token’s rendered image, r_i its within-collection scarcity, w_t a vector of market controls, $c(i)$ its collection and $\tau(t)$ its month. The specification is

$$\log p_{it} = x_i' \beta + \gamma r_i + w_t' \delta + \alpha_{c(i)} + \mu_{\tau(t)} + \varepsilon_{it}, \quad (2)$$

where α_c are collection effects and μ_τ month effects. The outcome is $\log p_{it}$. Sales below 5% of their collection’s trailing seven-day median are screened out as unlikely to be arm’s-length.

Three features of (2) carry the argument. The descriptors x_i are *token constant*: an image does not change between sales, so β is identified entirely from variation across tokens, never within one. The scarcity term r_i is likewise token constant and, as Section 2.4 noted, is a deterministic function of the same attribute layers that generate x_i . And α_c enters as a fixed effect rather than a random one, so no assumption about $\text{Cov}(\alpha_c, x_i)$ is required.

4.2 Within- versus between-collection identification

Tokens within a collection share a palette, a rendering convention and a subject, so much of the variation in x_i is variation *between* collections. A specification that pools the two estimates a weighted average of a within-collection premium and a between-collection one, and the literature’s practice of reporting a single coefficient does not distinguish them.

Three estimators separate the channels. The first is the crossed random-effects specification of the kind used in prior work, which pools. The second replaces α_c with collection dummies and identifies β from within-collection variation alone. The third is the correlated-random-effects specification of Mundlak (1978), which augments (2) with the collection means $\bar{x}_c$:

$$\log p_{it} = x_i' \beta_W + \bar{x}_{c(i)}' \theta + \gamma r_i + w_t' \delta + \mu_{\tau(t)} + \varepsilon_{it}. \quad (3)$$

Here β_W is the within-collection premium, $\beta_W + \theta$ the between-collection one, and θ their difference. The joint test $H_0: \theta = 0$ is exactly the orthogonality test that a pooled random-effects specification requires and does not report: under H_0 the two channels coincide and pooling is innocuous, and when it is rejected the pooled coefficient is a composite of two distinct objects.

4.3 Inference at the collection level

The descriptors vary at the level of the image and, in substantial part, at the level of the collection. The number of independent units is therefore not the number of sales but something closer to the number of collections, of which there are 26. Cluster-robust standard errors assume many clusters; with between five and thirty, the asymptotic approximation over-rejects, sometimes severely (Cameron et al., 2008; MacKinnon et al., 2023a, 2023b).

Inference is therefore conducted with a wild cluster bootstrap- t at the collection level, imposing the null, with Rademacher weights and 9,999 replications. Results are reported alongside four other error assumptions (independent, token-clustered, two-way token-by-date, and analytic collection-clustered), so that the consequence of each assumption is visible rather than asserted. Because 196 descriptors are tested, p -values are additionally corrected for multiplicity by Romano–Wolf stepdown on common bootstrap draws (Romano & Wolf, 2005; Clarke et al., 2020), and by Bonferroni, Holm and Benjamini–Hochberg for comparison. Where a descriptor’s identifying variation is concentrated in few collections, a Kish-style effective-cluster count is reported as a diagnostic, in the spirit of Carter et al. (2017).

4.4 What a token random effect can and cannot absorb

Prior specifications in this literature include a token random intercept and, on that basis, treat persistent token-level differences (including scarcity) as absorbed. The following result shows that the intercept cannot be relied upon for that purpose when the regressors are token constant.

Proposition. *Consider the mixed model $y = X\beta + Zu + \varepsilon$ with a token random intercept $u \sim \mathcal{N}(0, \sigma_u^2 I)$, $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 I)$, and token-constant regressors, so that $X = ZX_t$ for a matrix X_t of token-level characteristics. Then the estimated random effects satisfy*

$$X_t' \hat{u} = 0 \tag{4}$$

exactly, for any $\lambda = \sigma_\varepsilon^2 / \sigma_u^2 > 0$.

The result follows in three lines from Henderson’s mixed-model equations and is proved in Appendix A.1; it holds also for the crossed token-and-collection specification used in prior work. Its consequence is the reverse of the usual reading. The token intercept is *forced* orthogonal to the included token-constant regressors, so it cannot absorb any component of an omitted token-constant variable that is correlated with them. It absorbs only the orthogonal remainder. The correlated component has nowhere to go except into $\hat{\beta}$.

Two implications follow, and it is important to keep them distinct. First, the observation that included features explain almost none of the estimated token effects is not evidence that the design is clean: (4) guarantees that outcome mechanically, whatever the truth. Second, and this is the substantive point, a token-level random-effects specification with

correlated omitted scarcity will report *inflated* feature premia rather than attenuated ones. In a synthetic design with a true feature effect of exactly zero, a correlated omitted regressor, and the estimated token effect orthogonal to the features to machine precision, the estimator returns +3.03.

Scope. The proposition concerns designs with a token *random* effect. It does not apply to token fixed effects, and it does not apply to the collection-fixed-effects specification of (2), which is what the results below use. It therefore explains why a token-level random-effects design tends to *find* visual premia; it does not predict the null reported in Section 5, which comes from a different specification. The two results are complementary, and neither implies the other.

5 Results

The results are presented in the order the argument requires: first where the variation in the descriptors lies, then what survives inference at the level of that variation, then what happens once scarcity is controlled, then the same exercise across all 196 descriptors, then time variation and out-of-sample prediction.

5.1 Where the variation is

Table 6 decomposes each premium into within- and between-collection components.

Table 6: Pooled, within-collection and between-collection premia

Feature	Between share	Pooled (A)	Within (B)	Between (C)	Verdict
<i>Line Art Proportion Curved</i>	90.9%	+0.400	−0.062	+1.018	sign flip
<i>Line Art Average Thickness</i>	81.3%	−0.502	+0.012	−1.458	sign flip
<i>Edge Bounding Box Area</i>	80.3%	+0.905	+0.332	+2.289	collapses
<i>Edge Range X</i>	77.1%	−0.207	−0.176	−0.907	survives
<i>Edge Range Y</i>	72.2%	−0.012	−0.117	−0.144	small in baseline
<i>Color Lightness Distribution</i>	53.3%	−0.203	+0.014	−0.440	sign flip
<i>Color Palette Mean Delta-E</i>	45.2%	−0.070	+0.007	+0.313	sign flip
<i>Composition Focus Saturation</i>	43.8%	+0.126	+0.013	+0.191	collapses
<i>Composition Focus Lightness</i>	23.4%	−0.108	+0.006	−1.142	sign flip
<i>Color Hue Distribution</i>	20.8%	−0.102	+0.007	−0.617	sign flip

Joint Mundlak test, H_0 : within = between for all 13 features: $\chi^2(13) = 14,118$, $p < 10^{-16}$
Across the 13 features: 1 survives, 2 collapse, 7 sign-flip, 3 small in baseline

Notes: $N = 93,701$ sales; 26 collections; 24,129 tokens. Coefficients are per one pooled standard deviation of the regressor. Column A is the crossed random-effects specification used in this literature; column B is collection least-squares dummy variables with month fixed effects, the pure within contrast; column C is the Mundlak between estimate $\beta_W + \theta$ from (3). “Between share” is the proportion of the descriptor’s variance lying between collections. The “survives” verdict denotes analytic significance only; under a scarcity control, a wild cluster bootstrap and a multiplicity correction no descriptor survives (Tables 7–9). In the correlated-random-effects model the combined token-and-collection variance is identified but its split is not, so within estimates are taken from the assumption-free column B.

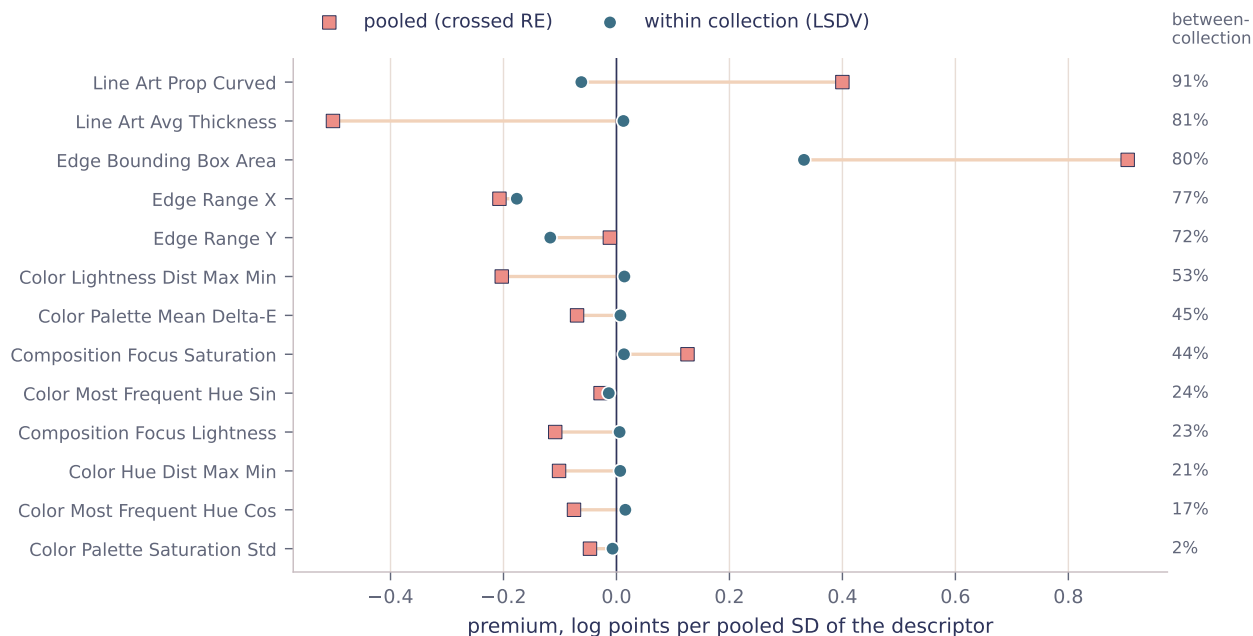


Figure 3: **The premia are between collections, not within them.** Each descriptor is shown twice: the pooled estimate from the crossed random-effects specification used in this literature, and the within-collection estimate from collection least-squares dummy variables with month fixed effects. The percentage at the right is the share of the descriptor's variance lying between collections. Descriptors are ordered by that share. Seven of the thirteen change sign under within-collection identification, including the two largest; the joint Mundlak test rejects equality of the two channels at $\chi^2(13) = 14,118$. Coefficients are per one pooled standard deviation of the descriptor. $N = 93,701$ sales, 26 collections. Position, not significance, is plotted, under collection-level inference none of these premia is identified (Figure 4).

Two things stand out. The five largest premia are between 72% and 91% *between* collections, so in a pooled specification they are principally measuring the collection. And under within-collection identification the estimates do not merely shrink: seven of the thirteen change sign, including the two largest. *Line Art Average Thickness* moves from -0.502 to $+0.012$ and *Line Art Proportion Curved* from $+0.400$ to -0.062 .

The joint test on the collection means rejects decisively, $\chi^2(13) = 14,118$. The within and between channels are not the same object, and a specification that pools them reports a composite of the two. This is the orthogonality condition that a pooled random-effects specification requires; it does not hold here.

A trait worth noting. *Composition Focus Saturation*, the descriptor that hedonic studies of digital art most often single out, is 43.8% between-collection and collapses from $+0.126$ to $+0.013$ under within-collection identification.

5.2 What survives inference at the collection level

Table 7 holds the point estimates fixed and varies only the error assumption.

Table 7: The same coefficients under five error assumptions

Feature	t-statistic under				Wild bootstrap p
	i.i.d.	token	two-way	collection	
<i>Line Art Proportion Curved</i>	+37.9	+17.0	+7.7	+1.23	0.376
<i>Line Art Average Thickness</i>	−49.6	−20.5	−10.7	−1.56	0.265
<i>Edge Bounding Box Area</i>	+30.7	+15.1	+5.4	+1.11	0.398
<i>Edge Range X</i>	−17.9	−9.2	−3.1	−0.71	0.576
<i>Color Lightness Distribution</i>	−35.0	−19.2	−7.5	−2.96	0.033
<i>Composition Focus Saturation</i>	+24.1	+13.2	+7.1	+1.56	0.208
<i>Color Hue Distribution</i>	−17.1	−9.9	−6.5	−2.02	0.081
<i>Color Palette Saturation Std</i>	−8.8	−2.1	−2.1	−1.34	0.023
<i>Significant at 5%, of 18 terms after any multiplicity correction</i>	15	15	12	2	0

Notes: $N = 93,701$; 24,129 tokens; 1,515 days; 26 collections. Ordinary least squares with month fixed effects; point estimates are identical across columns, only the error assumption changes. The wild cluster bootstrap- t imposes the null, uses Rademacher weights at the collection level and 9,999 replications, and is validated to four decimal places against the analytic estimator. The two raw bootstrap survivors do not survive multiplicity correction: across 18 simultaneous tests roughly one false positive is expected, and under Bonferroni, Holm or Benjamini–Hochberg none survives, the smallest Holm-adjusted p being 0.414. Both are colour-dispersion measures whose identifying variation is concentrated in one or two collections (Table 16).

The pattern is monotone in the plausibility of the assumption. Treating 93,701 sales as independent gives t -statistics between 9 and 50. Clustering by token (the correction for the fact that the same image is sold repeatedly) roughly halves them, and 15 of 18 terms remain significant, so the repeated-sales objection alone does not overturn the results. Clustering at the collection level, which is the level at which the descriptors actually vary, leaves two. Neither of those two survives correction for the fact that 18 hypotheses were tested at once. *Why the bootstrap rather than the sandwich.* The analytic cluster-robust estimator and the bootstrap disagree on which terms are significant, and the disagreement is systematic: the sandwich flags descriptors whose identifying variation sits in a handful of collections, where its asymptotic justification is weakest. Section 7.2 quantifies this per regressor.

5.3 Scarcity

Table 8 adds the scarcity measure that the design has so far omitted.

Scarcity is priced, and it is the only regressor in the analysis that clears collection-level inference. One within-collection standard deviation is worth +0.154 log points, roughly 16.7%, with a bootstrap p indistinguishable from zero.

Two things about this result are worth stating carefully. First, adding scarcity does not substantially move the visual coefficients: the largest absolute change is 0.012 in the crossed

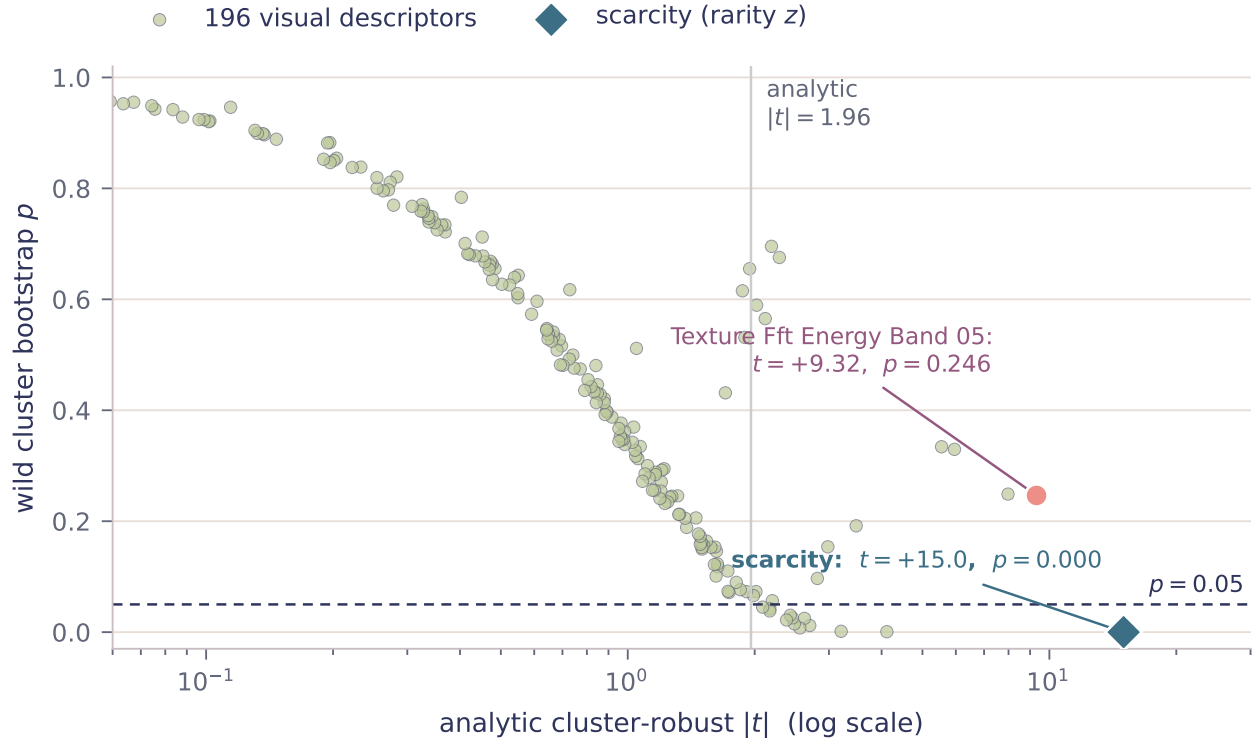


Figure 4: **At 26 clusters the analytic t -statistic does not order the evidence.** Each small circle is one of the 196 descriptors, entered one at a time in the within-collection specification with a scarcity control. The horizontal axis is its analytic cluster-robust $|t|$; the vertical axis is the wild cluster bootstrap p -value from the same regression, with the null imposed and $B = 9,999$ Rademacher draws. If the two agreed, every point right of $|t| = 1.96$ would sit below $p = 0.05$. Many do not. *Texture Fft Energy Band 05* carries $t = +9.32$ and a bootstrap p of 0.246; scarcity carries a similar t and a bootstrap p of 0.000. The difference is that scarcity varies within all 26 collections by construction, while the texture band's identifying variation is concentrated in a handful of them (Table 16, Panel B).

Table 8: The scarcity premium, and the visual descriptors beside it

Term	Coefficient	Analytic t	Bootstrap p
<i>Panel A. Scarcity</i>			
Scarcity, within-collection z (LSDV)	+0.1542	+15.00	< 0.0001
Scarcity, crossed random effects	+0.1589	+19.22	–
Scarcity, population denominators (18 colls.)	+0.1630	–	< 0.0001
Scarcity, excluding degenerate collections	+0.1657	–	< 0.0001
<i>Panel B. Visual descriptors, same specification, with scarcity</i>			
Edge Bounding Box Area	–	+1.90	0.1393
Edge Range Y	–	–1.95	0.1539
Edge Range X	–	–1.62	0.1907
Color Palette Saturation Std	–	–3.16	0.2025
Line Art Proportion Curved	–	–0.91	0.3873
Color Most Frequent Hue ($\cos$)	–	+1.76	0.0846

Notes: $N = 93,383$; 26 collections. Scarcity is the sum of inverse attribute frequencies, z -scored within collection, constructed from metadata collected for 23,997 tokens (99.4%) across all 26 collections. A one-standard-deviation increase corresponds to approximately +16.7%, computed as $\exp(0.1542) - 1$; the log coefficient is not itself a percentage. Bootstrap as in Table 7. Because scarcity is computed from sampled attribute frequencies, classical measurement error attenuates the coefficient toward zero, so +0.1542 is a lower bound; the population-denominator and degenerate-exclusion rows move in the direction attenuation predicts.

random-effects specification and 0.034 within, with no sign flips. The visual premia were therefore not principally mismeasured scarcity: the omitted-variable bias runs in the expected direction, but its magnitude is small.

Second, and more instructive, adding scarcity shrinks the *analytic* collection-clustered standard errors of the edge descriptors by roughly 28%, which makes several of them appear newly significant on the very statistic that Table 7 has already discredited at 26 clusters. Under the bootstrap, none of them is significant with scarcity or without it. A good control can manufacture apparent significance through the standard error alone, and the only defence is to conduct inference at the right level.

5.4 All 196 descriptors

The results so far concern 13 interpretable descriptors. Restricting attention to a screened subset invites the objection that the subset was chosen well or badly, so Table 9 instead tests all 196 descriptors individually, with no screening step of any kind.

No descriptor survives any multiplicity correction once scarcity is controlled. The uncorrected bootstrap count, twelve, is close to the ten expected by chance across 196 simultaneous tests. The null is a property of the descriptor set as a whole, not of any particular selection from it.

The comparison between the two columns is more informative than either alone. *Without* a scarcity control the descriptor set carries more signal than chance: seventeen clear the

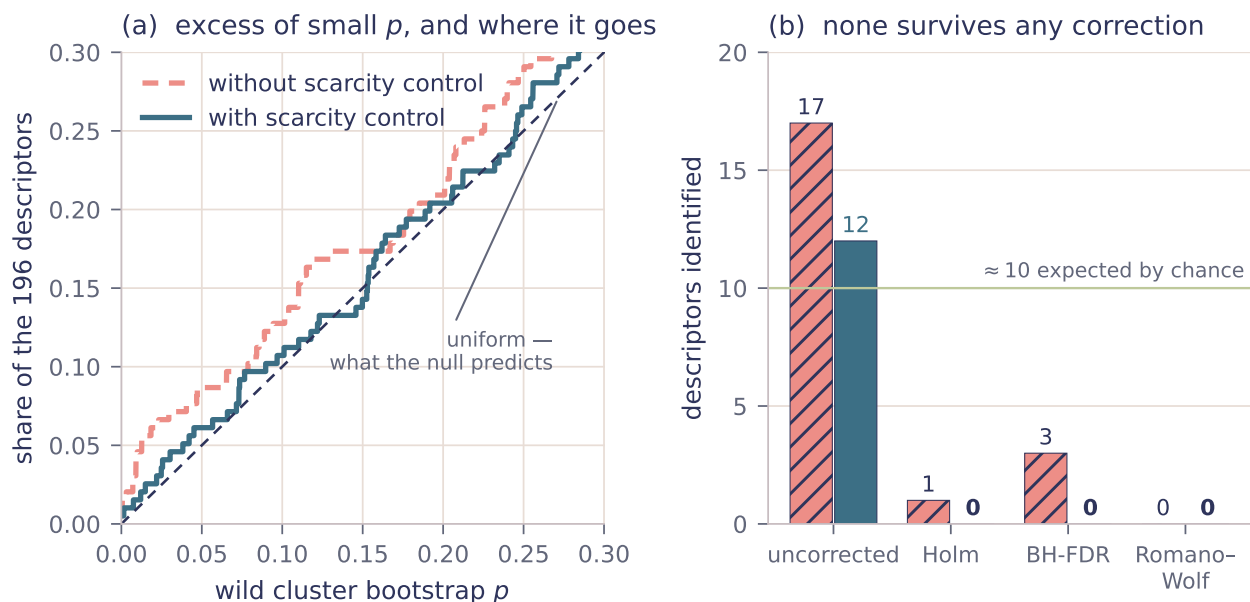


Figure 5: **What the scarcity control removes.** Panel (a) is the empirical distribution of the 196 bootstrap p -values, shown over the region where a real effect would appear. Under the null they should track the diagonal. Without a scarcity control the distribution lies visibly above it, there are more small p -values than chance produces. Adding the scarcity control returns the distribution to the diagonal. Panel (b) counts the descriptors called significant under each multiplicity correction. Without the scarcity control, 17 clear the uncorrected threshold, one survives Holm and three survive Benjamini–Hochberg. With it, 12 clear the uncorrected threshold, around the ten expected by chance, and *none* survives any correction. Romano–Wolf, which is the appropriate stepdown for 196 dependent tests on common draws, returns zero either way.

Table 9: All 196 visual descriptors, with and without a scarcity control

Descriptors significant at 5%	Without scarcity	With scarcity
Analytic cluster-robust t	–	26
Wild cluster bootstrap, uncorrected	17	12
<i>expected by chance</i>	≈ 10	≈ 10
Bonferroni / Holm	1	0
Benjamini–Hochberg	3	0
Romano–Wolf stepdown	0	0
<i>Best embedding, Romano–Wolf p (with scarcity)</i>		
Convolutional, 100 components		0.9972
Style/Gram, 50 components		0.3461
<i>Illustrative case: Color Hue Counts</i>		
Bootstrap p	< 0.0001	0.0015
Romano–Wolf p	0.3455	0.8041

Notes: Specification $\log p_{it} \sim \text{descriptor}_j + \text{controls} + \text{scarcity} + \text{collection fixed effects} + \text{month fixed effects}$, estimated once per descriptor, $N = 93,383$. Romano–Wolf stepdown is computed across all 196 hypotheses on common bootstrap draws. Bootstrap p -values are multiples of $1/9,999$, so “< 0.0001” denotes zero rejections among 9,999 draws rather than a literal zero. Of the 196 descriptors, 150 are black-box embeddings; none approaches significance. The embedding components were reduced by a principal-components step fitted separately within each collection, so a given component index denotes a different linear combination in each collection; their pricing performance should be read as a lower bound.

uncorrected bootstrap, one survives Bonferroni and three survive Benjamini–Hochberg. *With* scarcity controlled, none survives anything. *Color Hue Counts* illustrates the mechanism: its bootstrap p moves from below 0.0001 to 0.0015, and its Romano–Wolf p from 0.35 to 0.80. Within collection, the measurable visual signal is scarcity.

Why analytic statistics are set aside throughout. One texture descriptor carries an analytic t of +9.32 against a bootstrap p of 0.2464. Scarcity carries a comparable t and a bootstrap p below 0.0001. The difference is that scarcity varies within all 26 collections by construction while the texture band varies almost entirely between them.

5.5 Time variation

Table 10 allows collection premia, the scarcity premium and each of the thirteen descriptors to vary across periods.

Collection premia move substantially with the market, $\Delta R^2 = +0.167$. Because the test compares each collection against itself over time, it is not driven by sample composition. The scarcity premium also moves, but by two orders of magnitude less. No descriptor drifts on either partition.

Brand premia are time-varying; brand dispersion is not cyclical. Cross-collection dispersion appears to more than double between 2021H1 and 2021H2, but only 3 of 26 collections were active in 2021H1 and all 26 only from 2023H1. On a balanced panel the series is flat: 16 collections over 8 periods give a peak-to-trough ratio of 1.24, and 22 over 7 periods give 1.18.

Table 10: What moves with the market cycle

Interaction	Calendar half-years			Labelled cycles	
	F	q	ΔR^2 / boot p	ΔR^2	boot p
Collection $\times$ period (brand)	578.5	160	+0.1671 / 0.0010	+0.1789	< 0.0001
Scarcity $\times$ period	24.5	8	+0.0007 / 0.0142	+0.0008	0.0042
13 descriptors $\times$ period	–	8	0 of 13 significant	–	0 of 13

Notes: Sale-level, $N = 93,383$. Calendar half-years are the primary partition; the ten labelled cycles are retained for comparability but are defined from realised prices, so interacting descriptors with them conditions the partition on the outcome. The test statistic is the classical F from restricted-versus-unrestricted sums of squares with a wild cluster bootstrap null ($B = 4,999$), because a cluster-robust Wald statistic cannot invert more than approximately G restrictions and the brand test requires 160. Scarcity is interacted with period in every specification, so descriptor drift cannot be scarcity drift. Every descriptor’s Holm-adjusted p equals 1.0000 on both partitions.

The apparent doubling is collection entry.

One further illustration of the inference problem. *Edge Bounding Box Area* has $F = 364.5$ against a bootstrap p of 0.4201, an F of that size that is indistinguishable from noise once 26 clusters are respected.

5.6 Out-of-sample prediction

The tests so far ask whether a descriptor carries an identified premium. Table 11 asks the different question of whether the descriptors *predict*.

Table 11: Out-of-sample prediction of log price

Predictor block	Out-of-sample R^2		
	Random	By token	Held-out collection
Collection identity alone	0.4256	0.4251	–0.0426
All 196 descriptors	0.4717	0.4131	–0.0270
All 196 + collection	0.4754	0.4188	–0.0157
46 interpretable descriptors	0.4668	0.4167	–0.0561
150 embeddings	0.3166	0.1244	–0.0456
Scarcity alone	0.0275	0.0124	–0.0634

Notes: Gradient boosting, five folds, R^2 against the training mean, $N = 93,383$. The cross-validation design is part of the exercise: 94.5% of sales come from tokens that trade more than once and the descriptors are token constant, so a random split allows the model to memorise token price levels. The embedding block scores 0.3166 under a random split and 0.1244 when folds are grouped by token, an overstatement of 155%; the token-grouped column is the appropriate one. The held-out collection column should be read on the 46 interpretable descriptors only: the 150 embedding components were reduced within each collection, so across-collection generalisation is undefined for those blocks and their entries are reported for completeness. Collection identity bundles brand with launch timing, community, marketing and liquidity, so its predictive dominance does not establish a causal brand effect.

Collection identity alone predicts better than all 196 descriptors, and adding the descriptors

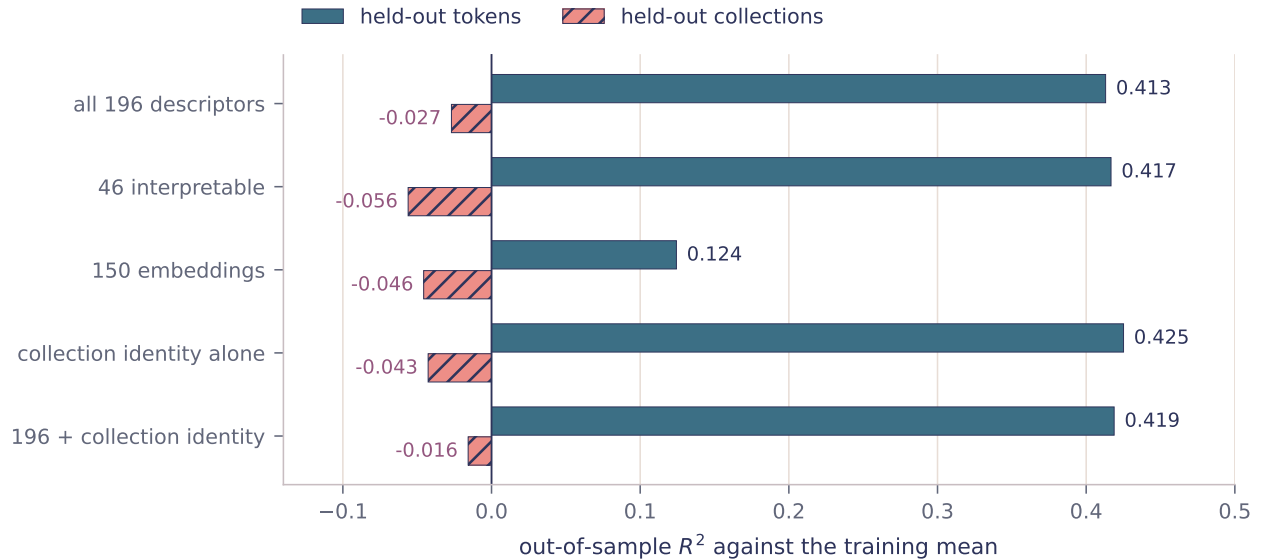


Figure 6: **Visual descriptors predict price within collections and not across them.** Out-of-sample R^2 against the training mean, gradient-boosted trees, five folds. Because 94.5% of sales come from tokens that sell more than once and the descriptors are token-constant, folds are grouped by token; a random split would overstate the embedding block by 155%. The solid teal bars hold out tokens from collections seen in training; the hatched coral bars hold out whole collections. Collection identity alone predicts held-out tokens slightly better than all 196 descriptors, and adding the descriptors on top of it changes little. Held out a whole collection, every block is worse than the training mean. The cross-collection reading is stated on the 46 interpretable descriptors, whose scale is common across collections; the embedding blocks are reported for completeness because their principal components were fitted separately within each collection.

on top of collection identity does not improve on collection identity alone. Held out an entire collection, the interpretable descriptors score -0.0561 : a model trained on 25 collections cannot price the 26th from its pixels, and does worse than predicting the training mean.

The information the descriptors carry about price is therefore information about *which collection a token belongs to*. That is a statement about where the predictive content lies, not about what causes value.

5.7 Comparison with prior findings

Table 12 sets these results against the prior literature mapped in Section 2. Three findings are confirmed, one does not carry over to prediction outside the collection, eight determinants have no prior estimate at all, and for a further group the comparison is unavailable rather than resolved.

Table 12: Comparison with prior findings.

Determinant	Prior evidence	Evidence here	Verdict
<i>Panel A. Directly comparable with prior work</i>			
Trait scarcity	Positive in every study that estimates it: Mekacher et al. (2022); Hubbard & LaFave (2026); Liu et al. (2025); Kong & Lin (2021) and Schaar & Kampakis (2022), the last two as reported in Kräussl & Tugnetti (2024)	+0.1542 log points, bootstrap $p < 0.0001$; the only regressor clearing collection-level inference	Confirms
Identity dominates the image	74.2% of out-of-sample variation with the image against 71.8% without, on 1.1m paintings (Aubry et al., 2023); collection past median price predicts more than half the variance (Nadini et al., 2021)	Collection identity alone 0.4251; all 196 descriptors 0.4131; both together 0.4188	Confirms, and sharpens: identity alone is better

continued on next page

Table 12, continued

Determinant	Prior evidence	Evidence here	Verdict
The image matters more where identity is unknown	The image’s contribution rises from +2.4 to +11.8 percentage points as characteristics are removed (Aubry et al., 2023); embeddings add value mainly for first-time sales (Mei et al., 2025)	Held out an entire collection, the interpretable descriptors score -0.056	Does not carry over: out of collection the descriptors do worse than the mean
Deep embeddings add little	Visual features alone reach $R_{adj}^2 \in [0, 0.08]$ (Nadini et al., 2021)	150 embeddings 0.1244 against 46 legible descriptors 0.4167; best embedding Romano–Wolf $p = 0.9972$	Confirms, and sharpens to black-box features specifically
<i>Panel B. No prior estimate exists</i>			
Edge spatial extent and framing (5)	Estimated on listing prices in two collections, where the descriptors that carry the relationship differ between them (Chen et al., 2023); never on realised prices with attributes held fixed	0 of 5 identified	First identified estimate
Style/Gram embeddings (50)	–	0 of 50 identified	New
Perceptual colour distance, ΔE (1)	–	Not identified	New
Spatial-frequency and LBP texture (15)	Prior work uses a single scalar complexity measure	0 of 15 identified	New
Sharpness (1)	A blurriness measure is among those related to listing prices in two collections (Chen et al., 2023)	Not identified	First identified estimate

continued on next page

Table 12, continued

Determinant	Prior evidence	Evidence here	Verdict
Brand premia are time-varying	No prior formal test	$F(160, 93128) = 578.5$, bootstrap $p = 0.0010$, $\Delta R^2 = +0.167$	New
Attributes are visually legible	No prior test	Median balanced accuracy 0.8708 against a chance rate of 0.0801	New
Visual features reach price only through the attribute vector	No prior test	Index priced at +0.0590 against measured scarcity; -0.0108 , $p = 0.2793$ against the full attribute vector	New
<i>Panel C. Comparison unavailable</i>			
Colour diversity; colour and line variance	Positive: Stepanova (2019); Borisov et al. (2022); Ju et al. (2019)	0 of 196 identified under within-collection identification	Neither confirmed nor contradicted
Saturation	+ Pownall & Graddy (2016); + Garay et al. (2022a); – Alsultan et al. (2024)	Not identified	Prior evidence itself conflicts
Lightness	– Pownall & Graddy (2016); + Garay et al. (2022a); – Alsultan et al. (2024)	Not identified	Prior evidence itself conflicts
Complexity	– Borisov et al. (2022); + Alsultan et al. (2024); + Ju et al. (2019); sign-unstable across six specifications in Fridgen et al. (2025)	Not identified	Prior evidence itself conflicts

Notes. **Panel C is not a set of contradictions.** With no descriptor identified under within-collection identification with collection-level inference and a scarcity control, there is no estimated effect to set against a prior one; the comparison is unavailable at this design rather than resolved against it. Nor would a contradiction be the natural reading if it were available: the studies concerned are on paintings, where each work is its own object and none of the three conditions of Section 2.4 (collection-level style, few effective clusters, and scarcity determined by the same attribute layers that generate the image) applies.

Where prior evidence conflicts with itself, this paper can supply a diagnosis rather than a verdict. The three determinants in Panel C with opposing signs across studies are exactly those with the lowest effective cluster counts in this sample, 1.6 to 1.7 of a nominal 26 (Table 16): a single study’s estimated sign for such a descriptor is decided by a handful of collections or artists, so sign reversals across samples are what one should expect. Panel B’s counts are the strict feature partition of Table 1.

6 What Visual Features Do Carry

Sections 5.1–5.4 established a null: no visual descriptor carries an identified within-collection premium. A null of that kind admits two readings. Either the images contain no economically relevant information, or they contain information that the pricing design cannot isolate. This section distinguishes them.

The tests below are conducted at the level of the *token* rather than the sale, on roughly 24,000 tokens. The small number of collections that constrains every pricing result in Section 5 therefore does not bind here, and these are the only well-powered estimates in the paper. The exercise is a predictive one, in the sense that Hernández-López and Wences (2024) use machine learning on a hedonic problem in housing: the question is how much of a quantity is recoverable, not what it is worth.

6.1 Predicting scarcity from pixels

Scarcity is computed from metadata: how uncommon a token’s attribute values are within its collection. The 196 descriptors are computed from pixels. They are independent measurements of the same object and nothing requires them to agree. Table 13 asks whether they do.

Pixels predict within-collection scarcity. The relationship is strongly non-linear (trees reach 0.1599 where a linear model reaches 0.0223, seven times less), which is why the one-descriptor- at-a-time linear tests of Section 5 could not have detected it. It is not an artefact of the pipeline: a shuffled target returns -0.0084 . And it holds in 22 of the 26 collections individually.

The block comparison locates it. *Texture* is the strongest single family and *composition* is the weakest by an order of magnitude. The five most important individual descriptors, by permutation importance, are image entropy (+0.0222), a local-binary-pattern texture measure (+0.0200), a spatial-frequency band (+0.0186), image energy (+0.0164) and line-art fractal dimension (+0.0098). Four of the five are measures of how much visual information an image contains.

Substantively: rarer tokens look busier. Whatever the pixels are recovering, it is complexity and texture rather than composition or focal subject.

It is not attribute counting. Rare tokens carry more attribute layers, and the layer count alone explains $R^2 = 0.033$. Residualising scarcity on the layer count and collection effects

Table 13: Predicting within-collection scarcity from pixels

	R^2	Spearman	Collections $R^2 > 0$
<i>Panel A. Design and model</i>			
Gradient-boosted trees, random token folds	+0.1599	+0.4347	22 of 26
on a within-collection rank target	+0.3261	+0.5309	–
Ridge, identical folds	+0.0223	+0.1760	15 of 26
Leave one collection out	−0.1114	+0.1871	13 of 26
Shuffled-target null	−0.0084	–	–
<i>Panel B. Descriptor block</i> (trees, random token folds)			
46 interpretable descriptors	+0.1638		
All 196	+0.1599		
Texture	+0.1087		
150 embeddings	+0.0831		
Edge	+0.0756		
Colour	+0.0728		
Image quality	+0.0523		
Line art	+0.0487		
Composition	+0.0050		

Notes: $n = 23,985$ tokens, 26 collections, five folds, no hyperparameter tuning, so all figures are lower bounds. Scarcity is z -scored *within* collection, so $E[y \mid \text{collection}] = 0$ exactly and a model can earn no R^2 from recognising which collection it is looking at; Panel A therefore cannot be restated as pixels encoding collection identity. The rank-target row replaces the z -score with a within-collection percentile, which is bounded and removes the skew of the raw measure. Block comparisons hold *in this benchmark only*: no block was separately tuned and they range from three columns to 150. The leave-one-collection-out row should be read on the interpretable blocks; the embedding components were reduced within each collection, so a component index denotes a different linear combination in each.

moves the result from +0.1781 to +0.1572: roughly 12% of the signal is layer counting and 88% is not.

6.2 Which attributes are visually legible

If pixels recover scarcity, the natural mechanism is that they recover the attributes from which scarcity is computed. Table 14 tests that directly, one collection–trait pair at a time.

Table 14: Which attributes can be read from the image

<i>Panel A. By attribute category</i>		
Attribute category	Collections	Median skill
Background	17	+0.9931
Head	7	+0.9321
Body	8	+0.9215
Hair	5	+0.8822
Clothing	7	+0.8804
Face	6	+0.8411
Type	5	+0.8234
Skin	7	+0.6041
Eyes	10	+0.3609
Mouth	9	+0.2483
<i>Panel B. By descriptor family, 96 detectable pairs</i>		
Convolutional embeddings	100 features	+0.8105
Style/Gram embeddings	50 features	+0.7969
Texture	15 features	+0.2938
Colour	14 features	+0.2616
Line art	6 features	+0.0863
Composition	3 features	+0.0663
Image quality	3 features	+0.0532
Edge	5 features	+0.0515
<i>Backgrounds are the exception</i>		
Colour alone		+0.909
Convolutional		+0.870
Style/Gram		+0.797

Notes: 138 collection–trait pairs across 25 collections; 118 in unflagged collections, which is the sample reported here. Skill is $(\text{balanced accuracy} - 1/K)/(1 - 1/K)$, so zero is chance and one is perfect. Median balanced accuracy on these pairs is 0.8708 against a median chance rate of 0.0801; 84 of 118 pairs reach a skill of 0.50 or better, and only 2 of 118 fail to beat their own label-permuted null, on which median skill is +0.0013. The primary classifier is linear for comparability across pairs, so these are lower bounds. Family comparisons hold *in this benchmark only*. Detectability is not explained by data shape: the largest rank correlation between skill and either the number of classes or the number of tokens used is 0.068 across all 138 pairs.

Attributes are visually legible, and unevenly so. Large regions are read almost perfectly: backgrounds at a skill of 0.99, and heads, bodies, hair and clothing above 0.88. Small facial

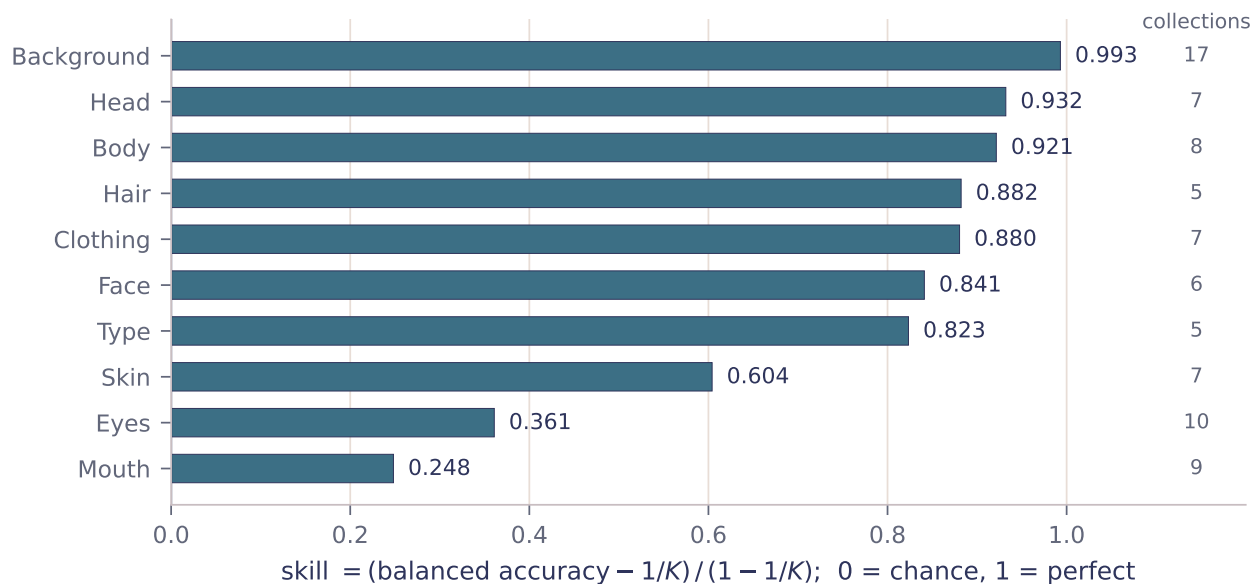


Figure 7: **Large regions are visually legible; faces are not.** For each (collection, attribute type) pair, a classifier is trained to recover the token’s attribute value from its image alone. Skill rescales balanced accuracy so that 0 is chance and 1 is perfect, which makes trait types with different numbers of classes comparable. Median skill across collections is plotted, restricted to attribute types appearing in at least five collections; the count at the right is the number of collections contributing. Background is read almost perfectly, and the parts of the image a viewer would look at first (eyes and mouth) are read worst. The permuted-label null is at most 0.006 for every trait type shown, so none of this is an artefact of class structure.

features are not: eyes at 0.36 and mouths at 0.25.

The split is coarse against fine, and it runs opposite to the attention of a human collector, for whom the face is the object of interest. It is also the reverse of the family ranking in Table 13: there the interpretable descriptors won, here the embeddings do, by a wide margin. The two tasks are different. Asking *how rare* a token is rewards a measure of how much is in the image, which the interpretable descriptors provide. Asking *which attribute* it carries is object recognition, which is what convolutional representations are built for.

One interpretable family wins outright. For backgrounds (large, flat regions of uniform colour), the fourteen colour descriptors alone reach +0.909, ahead of the hundred convolutional components at +0.870. It is the only attribute category where a named, interpretable family beats the embeddings.

6.3 Does the visual channel reach price?

Sections 6.1 and 6.2 establish that pixels carry information about a token’s attributes. Section 5 established that no descriptor carries an identified premium. Table 15 joins them, by constructing a visual index and asking what it is worth.

The index is an out-of-fold prediction of within-collection scarcity built from the 46 interpretable descriptors only, standardised and merged to the sale panel. It is, by construction, the part of the image that speaks to scarcity.

Table 15: Does the visual channel reach price beyond the attribute vector?

Specification	Coef.	t	Boot p	Var. retained
Visual index alone	+0.0998	–	< 0.0001	–
against measured scarcity	+0.0590	+4.29	< 0.0001	84.3%
against scarcity deciles	+0.0667	+5.24	0.0002	81.5%
against better-measured scarcity	+0.0582	+4.21	< 0.0001	83.6%
against the full attribute vector	–0.0108	–1.16	0.2793	19.6%
Placebo, shuffled target, linear control	–0.0054	–1.06	0.3261	99.3%
Placebo, shuffled target, attribute vector	–0.0034	–0.74	0.4819	61.6%

Notes: $N = 93,383$. All specifications include collection fixed effects, month fixed effects and market controls, with a wild cluster bootstrap at the collection level, $B = 4,999$. The final specification conditions on 7,342 collection-specific attribute indicators, absorbed by Frisch–Waugh–Lovell with a sparse solve. That absorption is exact by construction rather than approximate: a sum-of-inverse-frequency scarcity score is an additive function of the attribute indicators and therefore lies in their span, so no functional form is assumed. “Var. retained” is the share of the index’s variance surviving absorption. Measured scarcity alone enters at +0.1587 in this design.

The visual index is priced. Alone it is worth +0.0998 log points. Against the measured scarcity score it survives at +0.0590, so it is not merely a noisy copy of that score. It also survives a better-measured scarcity score built from full-population attribute frequencies, and a flexible control that enters scarcity as decile dummies rather than linearly. On this evidence the visual channel carries something the scalar scarcity measure does not.

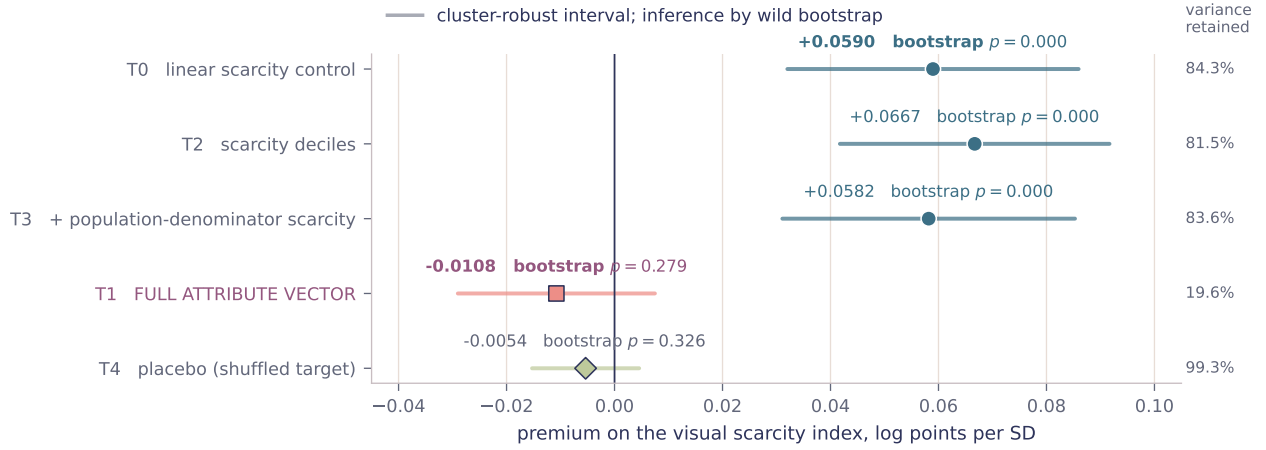


Figure 8: **The visual channel reaches price through the attribute vector, and no further.** The plotted coefficient is the premium on a visual scarcity index built out-of-fold from the 46 interpretable descriptors only. It survives a linear scarcity control (T0), a flexible decile control (T2) and a better-measured scarcity score built from full-population attribute counts (T3). It does not survive conditioning on the attribute vector itself (T1), which absorbs scarcity exactly rather than approximately, because the score $\sum_j 1/\text{freq}(j)$ is additive in the attribute indicators and so lies in their span. The null in T1 is not a power failure: 19.6% of the index’s variance is retained, and an effect of the T0 size would imply $t \approx +2.06$ against the -1.16 observed. A placebo index built on a shuffled target (T4) is dead in every specification. Inference is by wild cluster bootstrap at 26 clusters, $B = 4,999$; the bars show the cluster-robust interval for scale. The result is a conditional-independence statement, not a causal one.

Conditioning on the full attribute vector removes it entirely. The coefficient falls to -0.0108 , 18% of its size and with the sign reversed, at a bootstrap p of 0.28.

The null is mediation, not low power. Nearly a fifth of the index’s variance survives the absorption. If an effect of $+0.0590$ were genuine and independent of the attribute vector, that retained variance implies a t of approximately $+2.06$; the observed value is -1.16 . Both placebos are dead, so the construction does not manufacture significance.

What this identifies. The scalar scarcity score is a lossy summary of the attribute vector, which is why a visual index that recovers more of that structure beats the summary. But once the structure itself is conditioned on, the pixels add nothing. **Visual features price these tokens to the extent that they reveal which attributes a token carries, and no further.**

Two limits are worth stating. This is a conditional-independence result, not a causal one: visual features are not randomly assigned, and identifying a causal aesthetic premium would require exogenous variation in appearance holding the attribute vector fixed. And it does not establish whether attributes are priced *because of* how they look; it establishes that, given the attributes, appearance carries no further price information.

6.4 Do visual premia differ for rarer tokens?

A final possibility is that visual descriptors are priced, but only in part of the distribution. Each of the 13 interpretable descriptors was interacted with scarcity in three ways: continuously, and by splitting scarcity into terciles and quintiles. **None is significant in any design:** zero of 13 under the bootstrap in each case, with smallest bootstrap p -values of 0.2941, 0.0598 and 0.0528, and a largest incremental R^2 of 0.00034 across all 39 tests.

A suggestive but unidentified pattern. Three edge and framing descriptors rise monotonically with scarcity. *Edge Bounding Box Area* moves from $+0.0213$ in the most common third to $+0.0456$ in the middle and $+0.0942$ in the rarest, roughly four times larger, with a bootstrap p of 0.0598. The pattern is orderly, but across 13 simultaneous tests roughly one near-miss is expected, and the data cannot separate it from noise at this number of clusters. It is reported as suggestive and is not a finding.

A methodological by-product. That same descriptor carries a classical F of 39.13 with $p = 10^{-17}$, a cluster-robust Wald p of 0.0044, and a bootstrap p of 0.0598. The clustered Wald flags four of 13 descriptors in the tercile design and seven of 13 in the quintile design, where the bootstrap flags none. **Clustering alone is not sufficient at this number of clusters; the bootstrap is required.**

7 Robustness

Four checks. The first asks whether the null of Section 5 is an artefact of functional form. The second asks why analytic clustered inference and the bootstrap disagree, and supplies a

diagnostic that predicts where they will. The third varies the sample and the specification. The fourth asks whether more data would change anything.

7.1 Functional form

Every test in Section 5 enters each descriptor linearly. If the relationship between a descriptor and price were non-monotone (a premium at intermediate values, say, and a discount at both extremes), a linear term would recover nothing, and the null would be a null only for linear effects.

Panel A of Table 16 forecloses that. Eighteen descriptors were tested: the 13 interpretable ones and the five strongest from the full sweep by absolute t -statistic. Each was entered twice more, once as decile step dummies and once as a cubic B-spline added on top of the linear term, with the same wild cluster bootstrap and a Holm correction across descriptors.

Nothing survives in either flavour. One descriptor clears the uncorrected bootstrap under each, against roughly one expected by chance across 18 simultaneous tests, and none survives correction. The largest incremental R^2 from any non-linear term anywhere is 0.0008. There is no shape hiding behind the slope.

This forecloses non-linearity in the tested descriptors; it does not test interactions among them. Section 6.4 tests one family of interactions (with scarcity) and also finds nothing.

7.2 Why analytic clustered inference fails here

Table 7 showed the analytic cluster-robust estimator and the bootstrap disagreeing, and Section 6.4 showed the same disagreement in a different design. Panel B of Table 16 explains it, one number per regressor.

The nominal 26 collections are not 26 independent units for every regressor. Unequal collection sizes alone reduce the effective count to 21.1. Scarcity retains 19.7, because it varies within every collection by construction. The interpretable descriptors retain a median of 12.6, less than half.

At the bottom of the distribution the reduction is severe, and it is exactly there that the analytic estimator misleads. *Texture FFT Energy Band 05* draws effectively all of its identifying variation from three collections and has an effective count of 1.6, and it is the descriptor with an analytic t of +9.32 against a bootstrap p of 0.2464. The two descriptors that cleared the raw bootstrap in Table 7 have effective counts of 1.7.

The practical claim. Cluster-robust standard errors are the standard remedy, and here they are not enough: which regressors they fail for is predictable in advance from where the identifying variation sits.

Table 16: Functional form, and the effective number of clusters

<i>Panel A. Is the null an artefact of linearity?</i> (18 descriptors)		
	Decile steps	Cubic spline
Uncorrected bootstrap $p < 0.05$	1	1
<i>expected by chance</i>	≈ 1	≈ 1
Holm-corrected $p < 0.05$	0	0
Largest non-linear ΔR^2 observed		0.0008
<i>Panel B. Effective number of clusters</i> (nominal $G = 26$)		
	G_{eff}	% of 26
Size-based, from unequal collection sizes alone	21.1	81%
Scarcity	19.7	76%
13 interpretable descriptors, median	12.6	48%
<i>Color Palette Saturation Std</i>	1.7	7%
<i>Texture FFT Energy Band 05</i>	1.6	6%

Notes: Panel A: the 13 interpretable descriptors plus the five strongest from the 196-descriptor sweep, each entered as decile step dummies and, separately, as a cubic B-spline on top of the linear term; wild cluster bootstrap, $B = 4,999$, Holm correction across descriptors. Panel B reports a Kish-style concentration diagnostic in the spirit of Carter et al. (2017); it is not a verbatim implementation of their statistic, and the values are not literal counts of collections. Collection sizes range from 100 to 6,522 sales, a ratio of 65, which is what the size-based row reflects. The regressor-specific rows weight each collection by its share of the within-collection identifying variation in that regressor.

7.3 Sample and specification

Market controls. Seven specifications of the control set were estimated, varying which equity and crypto series enter and how non-trading days are handled. The chosen specification drops the S&P 500 return, which correlates 0.9524 with the NASDAQ return and carries a variance inflation factor of 10.78; the retained controls have factors between 1.02 and 3.66 (Section 3). Coefficients on the visual descriptors are stable across the seven.

The arm’s-length screen. Dropping sales below 5% of their collection’s trailing seven-day median removes 206 sales, 0.22% of the sample. Across the 21 image coefficients the largest resulting change is 0.0069 and there are no sign flips.

Duplicate feature rows. Two tokens carried four image-feature vectors each, inflating 12 real sales into 48. Collapsing them changes no coefficient by more than 0.000255 and no z -statistic by more than 0.020. The de-duplication key matters: collapsing on token, date and price alone would remove 6,972 rows, 7.4% of the sample, because dates are day-resolution and genuine repeat sales at an unchanged price are not duplicates. Only groups whose image features disagree are collapsed.

Scarcity measurement. Re-estimating with full-population attribute denominators for the 18 collections where they can be read on-chain gives +0.1630 rather than +0.1542; excluding the five collections with near-degenerate attribute distributions gives +0.1657. Both move in the direction that attenuation from measurement error predicts.

7.4 Would more data change the conclusions?

The sample covers roughly one token in ten. It is natural to ask whether the nulls of Section 5 are a power problem, and the question has a direct answer.

Modelling the full population of all tokens within these same 26 collections (roughly eleven times the current sample) moves **none of the five** tested descriptors from insignificance to significance. The median share of a descriptor’s standard error that additional tokens could ever remove is **13.6%**. The remainder is dispersion of the coefficient *across* collections, which more tokens within the same collections cannot touch.

That dispersion is itself informative. For eight of the nine terms examined, the standard deviation of the coefficient across collections exceeds the average coefficient’s own magnitude. For most visual descriptors there may be no single common premium waiting to be measured more precisely, which is consistent with the finding in Section 5.6 that a model trained on 25 collections cannot price the 26th. Scarcity is the exception: its cross-collection dispersion is well below its magnitude, which is why it is the one regressor that clears inference at 26 clusters.

What would change the conclusions is more collections, not more tokens. That is a different study, and a cheaper one: a few hundred tokens from each of many collections would be more informative than every token from these 26.

8 Concluding Remarks

What is a picture worth in a market where the picture can be copied freely and only the record of ownership is scarce? This paper measures 196 properties of 23,997 images, prices 93,701 sales of them, and arrives at an answer with two halves.

The first half is a null, and it is not a small one. No visual descriptor carries an identified within-collection premium. That holds across all 196 descriptors, under four different corrections for multiple testing once scarcity is controlled, under the Romano–Wolf stepdown whether or not it is, and under decile and spline specifications as well as linear ones. It is not a matter of statistical power: the full population of tokens within these same collections would move nothing, because what limits precision is dispersion of the coefficients *across* collections rather than noise within them.

Three separable things produce that result, and it is worth keeping them apart. Visual style is largely a collection-level property, so premia estimated from pooled variation measure the collection. The effective unit of inference is the collection, of which there are few, so statistics that assume many independent units overstate significance, often dramatically, as when a descriptor with an analytic t of +9.32 returns a bootstrap p of 0.246. And scarcity is determined by the same attribute layers that generate the image, so omitting it loads onto the visual coefficients rather than leaving them alone. None of the three is a criticism of any particular study. They are properties of the setting, and they are what a token-level hedonic

design in that setting has to survive.

Two things are priced. Scarcity is worth +0.154 log points per within-collection standard deviation, about 16.7%, and it is the only regressor in the analysis that clears inference at the collection level. Collection premia are large and move substantially with the market, while the scarcity premium moves with it only slightly and no visual descriptor moves with it at all.

The second half of the answer is where the pictures return. They are not uninformative. Within a collection, pixels predict which tokens are scarce, and they do it through texture and busyness rather than composition, rarer tokens simply contain more visual information. They also recover the attributes themselves, and unevenly: backgrounds are read almost perfectly, heads and bodies nearly so, while eyes and mouths are barely read at all. The parts of an image a machine finds legible are close to the opposite of the parts a collector looks at.

Joining the two halves gives the paper’s result. A visual index built to predict scarcity is priced, and survives both a better-measured scarcity score and a flexible one; it does not survive conditioning on the full attribute vector, where its coefficient falls to 18% of its size, reverses sign, and dies. The scalar scarcity score is a lossy summary of a token’s attributes; a visual index recovers more of that structure, which is why it beats the summary; and once the structure itself is held fixed, appearance adds nothing.

Visual features price these tokens to the extent that they reveal which attributes a token carries, and no further.

That is a statement about information, not about beauty. The images are doing real work (they are a measurement instrument for a token’s attribute structure, and a surprisingly good one), but they are not a separately compensated good. It is a conditional-independence result rather than a causal one: visual features are not randomly assigned, and identifying a causal aesthetic premium would require exogenous variation in appearance holding the attribute vector fixed, which observational data on these markets does not contain. Nor does the result establish whether attributes are priced *because of* how they look; it establishes that, given the attributes, appearance carries no further price information.

A further limitation concerns which sales are observed at all. Tokens trade when their holders choose to sell, and Huang and Goetzmann (2023) show for curated digital art that the resulting censoring is severe, with a large gap between observed and latent valuations at the market peak. The estimates here run on realised sales and are overwhelmingly repeat sales, so they inherit that concern. It would move the visual coefficients only if selection into trading correlated with the descriptors *conditional on* the attribute vector (a narrower requirement than selection correlating with price), but that is not testable with these data.

Three further limits bound what has been shown. The estimates apply to large, liquid generative collections on Ethereum, and generative collections are a particular kind of digital art: the finding that the attribute vector is what is priced may not carry to one-of-one works, where there is no attribute vector. The binding constraint throughout is the number of

collections, so a study spanning many more of them could establish whether any common visual premium exists at all, a question this design cannot settle either way. A few hundred tokens from each of a hundred collections would serve that purpose better than every token from twenty-six. And the feature set, though large, is a set of choices; a descriptor nobody has thought to compute cannot be shown to be unpriced.

What the design does settle is narrower and, we think, more useful. It shows what a token-level hedonic specification can and cannot identify when the objects being priced are draws from a common generative process, why specifications of that kind will tend to report visual premia that are not there, and what inference at a realistic number of independent units actually requires. Cluster-robust standard errors, the usual remedy, are not sufficient here, and which regressors they fail for can be predicted in advance from where the identifying variation sits.

References

- Alsultan, S., Kourtis, A., & Markellos, R. N. (2024). Can we price beauty? Aesthetics and digital art markets. *Economics Letters*, 235, 111572.
<https://doi.org/10.1016/j.econlet.2024.111572>
- Alternative.me. (n.d.). *Crypto Fear & Greed Index*. Retrieved September 23, 2025, from <https://alternative.me/crypto/fear-and-greed-index/>
- Ante, L. (2022). The non-fungible token (NFT) market and its relationship with Bitcoin and Ethereum. *FinTech*, 1(3), 216–224. <https://doi.org/10.3390/fintech1030017>
- Ashenfelter, O., & Graddy, K. (2003). Auctions and the price of art. *Journal of Economic Literature*, 41(3), 763–787. <https://doi.org/10.1257/002205103322436188>
- Aubry, M., Kräussl, R., Manso, G., & Spaenjers, C. (2023). Biased auctioneers. *The Journal of Finance*, 78(2), 795–833. <https://doi.org/10.1111/jofi.13203>
- Belousov, D., & Yanovich, Y. (2025). *DIT: Dimension reduction view on optimal NFT rarity meters* (arXiv preprint arXiv:2508.12671). <https://arxiv.org/abs/2508.12671>
- Borisov, M., Kolycheva, V., Semenov, A., & Grigoriev, D. (2022). The influence of color on prices of abstract paintings. *arXiv preprint arXiv:2206.04013*.
- Brahmana, R. K., Yeap, X. W., & Lean, H. H. (2025). Time–frequency connectedness among NFT assets. *Computational Economics*, 66(6), 4987–5013.
<https://doi.org/10.1007/s10614-025-10878-4>
- Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3), 414–427.
<https://doi.org/10.1162/rest.90.3.414>
- Carter, A. V., Schnepel, K. T., & Steigerwald, D. G. (2017). Asymptotic behavior of a *t*-test robust to cluster heterogeneity. *The Review of Economics and Statistics*, 99(4).
https://doi.org/10.1162/REST_a_00639

- Chanel, O., Gérard-Varet, L.-A., & Ginsburgh, V. (1996). The relevance of hedonic price indices. *Journal of Cultural Economics*, 20(1), 1–24. <https://doi.org/10.1007/BF00148268>
- Chen, J., & Roth, J. (2024). Logs with zeros? Some problems and solutions. *The Quarterly Journal of Economics*, 139(2), 891–936. <https://doi.org/10.1093/qje/qjad054>
- Chen, Y., Ye, Y., & Zeng, W. (2023). 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 (VINCI 2023)* (pp. 1–8). ACM. <https://doi.org/10.1145/3615522.3615545>
- Clarke, D., Romano, J. P., & Wolf, M. (2020). The Romano–Wolf multiple-hypothesis correction in Stata. *The Stata Journal*, 20(4), 812–843. <https://doi.org/10.1177/1536867X20976314>
- Cui, R., Deng, Y., Fu, B., & Xi, L. (2023). Exploring the price influences of PFP artworks in NFT based on regression analysis. In G. Vilas Bhau et al. (Eds.), *MSEA 2022, ACSR* (Vol. 101, pp. 390–396). https://doi.org/10.2991/978-94-6463-042-8_57
- Dowling, M. (2022a). Fertile LAND: Pricing non-fungible tokens. *Finance Research Letters*, 44, 102096. <https://doi.org/10.1016/j.frl.2021.102096>
- Dowling, M. (2022b). Is non-fungible token pricing driven by cryptocurrencies? *Finance Research Letters*, 44, 102097. <https://doi.org/10.1016/j.frl.2021.102097>
- Federal Reserve Bank of St. Louis. (n.d.). *FRED, Federal Reserve Economic Data*. Retrieved September 23, 2025, from <https://fred.stlouisfed.org/>
- Fridgen, G., Kräussl, R., Papageorgiou, O., & Tugnetti, A. (2025). Pricing dynamics and herding behaviour of NFTs. *European Financial Management*, 31(2), 670–710. <https://doi.org/10.1111/eufm.12506>
- Garay, U., Pérez, E., Casanova, J., & Kratochvil, M. (2022a). Color intensity, luminosity, contrast and art prices: The case of Jean-Michel Basquiat. *Academia Revista Latinoamericana de Administración*, 35(3), 303–328. <https://doi.org/10.1108/ARLA-05-2021-0110>
- Garay, U., Puggioni, G., Molina, G., & ter Horst, E. (2022b). A Bayesian dynamic hedonic regression model for art prices. *Journal of Business Research*, 151, 310–323. <https://doi.org/10.1016/j.jbusres.2022.06.055>
- Hernández-López, E., & Wences, G. (2024). A spatial regression approach in property valuation using machine learning. *Computational Economics*. Advance online publication. <https://doi.org/10.1007/s10614-024-10738-7>
- Holub, H. W., Hutter, M., & Tappeiner, G. (1993). Light and shadow in art price computation. *Journal of Cultural Economics*, 17, 49–69. <https://doi.org/10.1007/BF00820767>
- Huang, D., & Goetzmann, W. N. (2023). *Selection-neglect in the NFT bubble* (NBER Working Paper No. 31498). National Bureau of Economic Research. <https://www.nber.org/papers/w31498>
- Hubbard, T. P., & LaFave, D. (2026). Price differentials and skin tone in digital art. *Journal of Cultural Economics*, 50(1), 225–251. <https://doi.org/10.1007/s10824-025-09570-0>
- Ju, L., Tu, Z., & Xue, C. (2019). *Art pricing with computer graphic techniques* (Peking University

- HSBC Business School working paper; arXiv:1910.03800).
<https://arxiv.org/abs/1910.03800>
- Kang, H.-J., & Lee, S.-G. (2025). Market phases and price discovery in NFTs: A deep learning approach to digital asset valuation. *Journal of Theoretical and Applied Electronic Commerce Research*, 20, 64. <https://doi.org/10.3390/jtaer20020064>
- Kräussl, R., & Tugnetti, A. (2024). Non-fungible tokens (NFTs): A review of pricing determinants, applications and opportunities. *Journal of Economic Surveys*, 38(2), 555–574.
<https://doi.org/10.1111/joes.12597>
- Li, M., Tian, Y., Wang, J., Yao, H., Wei, Y., & Wang, S. (2026). How do traditional financial markets affect the NFT market? Insights into assets and sentiment impact. *Computational Economics*. Advance online publication. <https://doi.org/10.1007/s10614-026-11308-9>
- Liu, Y., Huangfu, B., Qiao, Z., & Zhao, X. (2025). Rarity, the WTA–WTP disparity, and price adjustments in the NFT market. *International Journal of Research in Marketing*, 42(3), 594–609. <https://doi.org/10.1016/j.ijresmar.2024.10.003>
- Ma, M. X., Noussair, C. N., & Renneboog, L. (2022). Colors, emotions, and the auction value of paintings. *European Economic Review*, 142, 104004.
<https://doi.org/10.1016/j.euroecorev.2021.104004>
- MacKinnon, J. G., Nielsen, M. Ø., & Webb, M. D. (2023a). Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics*, 232(2), 272–299.
<https://doi.org/10.1016/j.jeconom.2022.04.001>
- MacKinnon, J. G., Nielsen, M. Ø., & Webb, M. D. (2023b). Fast and reliable jackknife and bootstrap methods for cluster-robust inference. *Journal of Applied Econometrics*, 38(5), 671–694. <https://doi.org/10.1002/jae.2969>
- Mei, J., & Moses, M. (2002). Art as an investment and the underperformance of masterpieces. *American Economic Review*, 92(5), 1656–1668.
<https://doi.org/10.1257/000282802762024719>
- Mei, J., Moses, M., Wälty, J., & Yang, Y. (2025). *Deep learning for art market valuation* (arXiv preprint arXiv:2512.23078). <https://arxiv.org/abs/2512.23078>
- Mekacher, A., Bracci, A., Nadini, M., Martino, M., Alessandretti, L., Aiello, L. M., & Baronchelli, A. (2022). Heterogeneous rarity patterns drive price dynamics in NFT collections. *Scientific Reports*, 12, 13890. <https://doi.org/10.1038/s41598-022-17922-5>
- Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, 46(1), 69–85. <https://doi.org/10.2307/1913646>
- Nadini, M., Alessandretti, L., Di Giacinto, F., Martino, M., Aiello, L. M., & Baronchelli, A. (2021). Mapping the NFT revolution: Market trends, trade networks, and visual features. *Scientific Reports*, 11, 20902. <https://doi.org/10.1038/s41598-021-00053-8>
- Özdemir, O., & Kumar, A. S. (2024). Dynamic efficiency and herd behavior during pre- and post-COVID-19 in the NFT market: Evidence from multifractal analysis. *Computational Economics*, 63(3), 1255–1279. <https://doi.org/10.1007/s10614-023-10522-z>
- Pala, M., & Sefer, E. (2024). NFT price and sales characteristics prediction by transfer learning of

- visual attributes. *The Journal of Finance and Data Science*, 10, 100148.
<https://doi.org/10.1016/j.jfds.2024.100148>
- Pownall, R. A. J., & Graddy, K. (2016). Pricing color intensity and lightness in contemporary art auctions. *Research in Economics*, 70(3), 412–420.
<https://doi.org/10.1016/j.rie.2016.06.007>
- Renneboog, L., & Spaenjers, C. (2013). Buying beauty: On prices and returns in the art market. *Management Science*, 59(1), 36–53. <https://doi.org/10.1287/mnsc.1120.1580>
- Romano, J. P., & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4), 1237–1282. <https://doi.org/10.1111/j.1468-0262.2005.00615.x>
- Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82(1), 34–55. <https://doi.org/10.1086/260169>
- Stepanova, E. (2019). The impact of color palettes on the prices of paintings. *Empirical Economics*, 56(2), 755–773. <https://doi.org/10.1007/s00181-017-1413-4>
- Tang, M., Feng, X., & Chen, W. (2023). Exploring the NFT market on Ethereum: A comprehensive analysis and daily volume forecasting. *Connection Science*, 35(1), 2286912.
<https://doi.org/10.1080/09540091.2023.2286912>
- The Block. (2025). *Ethereum NFT marketplace monthly volume (wash-trading filtered)*.
<https://www.theblock.co/data/nft-non-fungible-tokens/marketplaces/nft-marketplace-monthly-volume>
- Yahoo. (n.d.). *Yahoo Finance*. Retrieved September 23, 2025, from <https://finance.yahoo.com/>
- Ziemke, V., Estermann, B., Wattenhofer, R., & Wang, Y. (2023). What determines the price of NFTs? In *2023 IEEE 29th International Conference on Parallel and Distributed Systems (ICPADS)* (pp. 1562–1568). IEEE. <https://doi.org/10.1109/ICPADS60453.2023.00220>

Statements and Declarations

Funding. No funding was received for conducting this study, and no support of any kind was received from any organisation for the submitted work.

Competing Interests. There are no relevant financial or non-financial interests to disclose.

Data Availability. The data that support the findings of this study are available from the author upon reasonable request.

A Appendix

A.1 Proof of the orthogonality proposition

Proposition. Consider the mixed model $y = X\beta + W\delta + Zu + \varepsilon$, where Z is the $n \times J$ token design matrix, X collects token-constant regressors so that $X = ZX_t$ for some $J \times k$ matrix X_t , W collects regressors that vary within token, $u \sim (0, \sigma_u^2 I_J)$, $\varepsilon \sim (0, \sigma_\varepsilon^2 I_n)$ and $\lambda = \sigma_\varepsilon^2 / \sigma_u^2 \in (0, \infty)$. Then the BLUP $\hat{u}$ from Henderson’s mixed model equations satisfies $X_t' \hat{u} = 0$ exactly.

Proof. Henderson’s equations give, for the u block,

$$Z'(y - X\hat{\beta} - W\hat{\delta}) = (Z'Z + \lambda I) \hat{u}, \quad (5)$$

and, for the fixed-effect block,

$$X'(y - X\hat{\beta} - W\hat{\delta} - Z\hat{u}) = 0. \quad (6)$$

Because X is token-constant, $X = ZX_t$ and hence $X' = X_t'Z'$. Substituting into (6),

$$X_t'Z'(y - X\hat{\beta} - W\hat{\delta}) = X_t'Z'Z \hat{u}. \quad (7)$$

The left-hand side of (7) is X_t' times the left-hand side of (5). Substituting (5) into (7),

$$X_t'(Z'Z + \lambda I) \hat{u} = X_t'Z'Z \hat{u} \implies \lambda X_t' \hat{u} = 0.$$

Since $\lambda > 0$, $X_t' \hat{u} = 0$. □

W drops out of the argument: it enters (5) and (7) identically, so the identity holds whatever within-token regressors are included. The argument goes through unchanged for the crossed token-and-collection specification used in prior work, where the collection term appears in both equations and cancels in the same way.

Two corollaries. First, regressing $\hat{u}$ on X_t returns a coefficient vector of exactly zero and an R^2 of exactly zero. The observation that included features explain almost none of the estimated token effects is therefore imposed by the estimator rather than discovered in the data, and would hold identically if scarcity mattered enormously. Second, an omitted token-constant regressor correlated with the included features cannot be absorbed by the token effect; the correlated component loads onto $\hat{\beta}$.

Numerical verification. On a synthetic unbalanced panel of $J = 400$ tokens and 1,803 sales with $k = 5$ token-constant features whose true effect is exactly zero, and an omitted token-constant regressor correlated with them, solving the mixed model equations directly gives $\max |X_t' \hat{u}|$ between 1.7×10^{-9} and 1.7×10^{-13} as λ ranges over $\{0.01, 0.1, 1, 10, 100\}$,

while $\max |\hat{\beta}|$ ranges over 3.02 to 3.05 against a true value of zero. At $\lambda = 1$ the estimator returns +3.03. In the crossed token-and-collection case $\max |X'_t \hat{u}| = 1.7 \times 10^{-11}$, so the identity still holds exactly. The script that produces these figures is provided with the submission.

A.2 The full descriptor set

Table 17: **The full set of 196 image descriptors and their results in the 196-feature sweep.** Every image descriptor constructed for this study is listed. None were screened out before estimation and none is degenerate on the estimation sample. Each row is a separate regression of the form $\log(\text{price}) \sim \text{descriptor}_j + \text{controls} + \text{rarity} + \text{collection FE} + \text{month FE}$, estimated on $N = 93,383$ sales of 23,985 tokens in 26 collections. Descriptors are standardised to unit variance on the estimation sample, so *Coef.* is the change in log price per one-standard-deviation change in the descriptor. p_{boot} is a wild cluster bootstrap- t p -value (Rademacher weights, null imposed, $G = 26$ clusters, $B = 9,999$ replications), so its resolution floor is $1/9,999$ and an entry of <0.0001 means no bootstrap draw exceeded the observed statistic. p_{RW} is the Romano–Wolf stepdown p -value, which corrects for testing all 196 descriptors jointly. The right-hand pair repeats the exercise with the trait-rarity control removed. *Tax.* gives the row of the literature taxonomy (Table 1) into which the descriptor falls. **In the specification that controls for trait rarity, none of the 196 descriptors survives any multiplicity correction:** across all 196 the smallest Romano–Wolf p -value is 0.2673, the smallest Holm-adjusted p -value is 0.1568, and the smallest Benjamini–Hochberg q -value is 0.1470. Individual descriptors therefore carry no identified premium once the multiplicity of the search is accounted for. Without the rarity control the picture is not identical: 3 descriptors reach Benjamini–Hochberg $q < 0.10$ on the raw bootstrap p -values, although none survives the Romano–Wolf correction, whose smallest value without rarity is 0.3415. That contrast between the two right-hand pairs is the object of interest, not a nuisance. Analytic collection-clustered t -statistics are reported in the replication file but are deliberately omitted here: with only 26 clusters they are not the basis for inference anywhere in this paper.

Feature	Family	Tax.	With rarity control			Without rarity control	
			Coef.	p_{boot}	p_{RW}	p_{boot}	p_{RW}
Panel A. Legible descriptors (46)							
COLOR_HUE_COUNTS	COLOR	A4	0.0549	0.0015	0.8041	<0.0001	0.3455
COLOR_HUE_DIST_MAX_MIN	COLOR	A5	0.0056	0.3617	1.0000	0.1670	1.0000
COLOR_LIGHTNESS_DIST_MAX_MIN	COLOR	A5	0.0110	0.0735	1.0000	0.0653	1.0000

continued on next page

Table 17, continued from previous page.

Feature	Family	Tax.	With rarity control			Without rarity control	
			Coef.	p_{boot}	p_{RW}	p_{boot}	p_{RW}
COLOR_MEAN_HUE	COLOR	A1	-0.0075	0.3347	1.0000	0.3874	1.0000
COLOR_MEAN_LIGHTNESS	COLOR	A3	-0.0096	0.2781	1.0000	0.1793	1.0000
COLOR_MEAN_SATURATION	COLOR	A2	-0.0283	0.0766	1.0000	0.1098	1.0000
COLOR_MOST_FREQUENT_HUE	COLOR	A1	-0.0036	0.7007	1.0000	0.6465	1.0000
COLOR_MOST_FREQUENT_HUE_PIXEL_RATIO	COLOR	A1	-0.0297	0.1178	1.0000	0.0295	0.9998
COLOR_PALETTE_COLORFULNESS	COLOR	A4	-0.0058	0.6174	1.0000	0.7748	1.0000
COLOR_PALETTE_MEAN_DELTA_E	COLOR	A6	-0.0001	0.9946	1.0000	0.8748	1.0000
COLOR_PALETTE_SATURATION_MEAN	COLOR	A2	-0.0011	0.8819	1.0000	0.9526	1.0000
COLOR_PALETTE_SATURATION_STD	COLOR	A5	-0.0063	0.0968	0.9609	0.0856	0.9456
COLOR_PALETTE_VALUE_MEAN	COLOR	A3	-0.0044	0.5085	1.0000	0.9746	1.0000
COLOR_PALETTE_VALUE_STD	COLOR	A5	-0.0069	0.3438	1.0000	0.1101	1.0000
COMPOSITION_FOCUS_HUE	COMPOSITION	B1	-0.0035	0.6103	1.0000	0.8907	1.0000
COMPOSITION_FOCUS_LIGHTNESS	COMPOSITION	B1	0.0088	0.1011	1.0000	0.2239	1.0000
COMPOSITION_FOCUS_SATURATION	COMPOSITION	B1	0.0169	0.2945	1.0000	0.4073	1.0000
EDGE_AREA	EDGE	B2	0.0101	0.7839	1.0000	0.3895	1.0000
EDGE_BOUNDING_BOX_AREA	EDGE	B2	0.0558	0.0303	0.9990	0.0031	0.7405
EDGE_MEAN_ENERGY_RATIO	EDGE	B2	-0.0203	0.2557	1.0000	0.0126	1.0000
EDGE_RANGE_X	EDGE	B2	0.0438	0.0382	1.0000	0.0079	0.9448
EDGE_RANGE_Y	EDGE	B2	0.0275	0.0715	1.0000	0.0090	0.9978
LINE_ART_AVG_ENDPOINTS_PER_SEGMENT	LINE_ART	B3	-0.0047	0.5116	1.0000	0.2131	1.0000
LINE_ART_AVG_LENGTH	LINE_ART	B3	-0.0058	0.3696	1.0000	0.0468	0.9456
LINE_ART_AVG_THICKNESS	LINE_ART	B3	0.0108	0.6641	1.0000	0.4352	1.0000
LINE_ART_FRACTAL_DIMENSION	LINE_ART	B3	0.0150	0.8825	1.0000	0.8601	1.0000
LINE_ART_PROP_CURVED	LINE_ART	B3	-0.0364	0.1230	1.0000	0.0091	0.8032
LINE_ART_PROP_STRAIGHT	LINE_ART	B3	0.0108	0.5965	1.0000	0.0944	1.0000
TEXTURE_FFT_ENERGY_BAND_01	TEXTURE	C2	0.0076	0.1917	0.6644	0.2009	0.7405
TEXTURE_FFT_ENERGY_BAND_02	TEXTURE	C2	0.0122	0.3342	0.3461	0.3331	0.3455
TEXTURE_FFT_ENERGY_BAND_03	TEXTURE	C2	0.0122	0.3295	0.3357	0.3265	0.3415
TEXTURE_FFT_ENERGY_BAND_04	TEXTURE	C2	0.0129	0.2489	0.3322	0.2788	0.3415
TEXTURE_FFT_ENERGY_BAND_05	TEXTURE	C2	0.0109	0.2464	0.2673	0.2682	0.3415
TEXTURE_LBP_00	TEXTURE	C2	0.0073	0.5653	1.0000	0.4953	1.0000
TEXTURE_LBP_01	TEXTURE	C2	0.0125	0.6755	1.0000	0.6025	1.0000
TEXTURE_LBP_02	TEXTURE	C2	0.0067	0.5894	1.0000	0.4633	1.0000
TEXTURE_LBP_03	TEXTURE	C2	0.0046	0.2456	1.0000	0.1754	1.0000
TEXTURE_LBP_04	TEXTURE	C2	0.0126	0.6154	1.0000	0.5050	1.0000
TEXTURE_LBP_05	TEXTURE	C2	0.0065	0.5314	1.0000	0.4155	1.0000
TEXTURE_LBP_06	TEXTURE	C2	0.0051	0.4314	1.0000	0.3771	1.0000
TEXTURE_LBP_07	TEXTURE	C2	0.0105	0.6552	1.0000	0.5386	1.0000
TEXTURE_LBP_08	TEXTURE	C2	0.0087	0.6955	1.0000	0.5733	1.0000
TEXTURE_LBP_09	TEXTURE	C2	0.0158	0.1538	0.9157	0.2038	0.9992
IMAGE_QUALITY_ENERGY	IMAGE_QUALITY	C1	-0.0758	0.0421	1.0000	0.0099	0.9703
IMAGE_QUALITY_ENTROPY	IMAGE_QUALITY	C1	0.0961	0.0075	0.9944	0.0006	0.7405
IMAGE_QUALITY_SHARPNESS	IMAGE_QUALITY	C3	0.0107	0.6434	1.0000	0.4395	1.0000
Panel B. Learned embeddings (150)							
CNN_PCA_001	CNN	D1	-0.0073	0.2707	1.0000	0.3510	1.0000
CNN_PCA_002	CNN	D1	-0.0038	0.5287	1.0000	0.2409	1.0000
CNN_PCA_003	CNN	D1	-0.0035	0.6784	1.0000	0.7392	1.0000
CNN_PCA_004	CNN	D1	0.0062	0.4426	1.0000	0.9629	1.0000
CNN_PCA_005	CNN	D1	0.0061	0.3672	1.0000	0.4519	1.0000
CNN_PCA_006	CNN	D1	-0.0030	0.6616	1.0000	0.7512	1.0000
CNN_PCA_007	CNN	D1	0.0011	0.8886	1.0000	0.8921	1.0000
CNN_PCA_008	CNN	D1	-0.0057	0.4306	1.0000	0.6522	1.0000

continued on next page

Table 17, continued from previous page.

Feature	Family	Tax.	With rarity control			Without rarity control	
			Coef.	p_{boot}	p_{RW}	p_{boot}	p_{RW}
CNN_PCA_009	CNN	D1	-0.0023	0.7590	1.0000	0.8785	1.0000
CNN_PCA_010	CNN	D1	0.0095	0.1569	1.0000	0.2068	1.0000
CNN_PCA_011	CNN	D1	0.0176	0.0567	1.0000	0.0787	1.0000
CNN_PCA_012	CNN	D1	-0.0046	0.5445	1.0000	0.4670	1.0000
CNN_PCA_013	CNN	D1	0.0064	0.4284	1.0000	0.3942	1.0000
CNN_PCA_014	CNN	D1	0.0009	0.9046	1.0000	0.9284	1.0000
CNN_PCA_015	CNN	D1	-0.0099	0.0256	0.9982	0.0472	0.9999
CNN_PCA_016	CNN	D1	-0.0034	0.4744	1.0000	0.0837	1.0000
CNN_PCA_017	CNN	D1	0.0050	0.3876	1.0000	0.3355	1.0000
CNN_PCA_018	CNN	D1	-0.0049	0.5277	1.0000	0.4968	1.0000
CNN_PCA_019	CNN	D1	0.0051	0.4206	1.0000	0.2674	1.0000
CNN_PCA_020	CNN	D1	0.0079	0.2544	1.0000	0.4444	1.0000
CNN_PCA_021	CNN	D1	0.0035	0.5371	1.0000	0.5291	1.0000
CNN_PCA_022	CNN	D1	0.0048	0.3127	1.0000	0.5044	1.0000
CNN_PCA_023	CNN	D1	0.0094	0.1771	1.0000	0.4759	1.0000
CNN_PCA_024	CNN	D1	0.0061	0.3278	1.0000	0.4585	1.0000
CNN_PCA_025	CNN	D1	0.0024	0.7342	1.0000	0.7643	1.0000
CNN_PCA_026	CNN	D1	0.0019	0.7253	1.0000	0.7375	1.0000
CNN_PCA_027	CNN	D1	-0.0016	0.7381	1.0000	0.9905	1.0000
CNN_PCA_028	CNN	D1	0.0094	0.0219	0.9995	0.0232	0.9985
CNN_PCA_029	CNN	D1	-0.0004	0.9216	1.0000	0.3754	1.0000
CNN_PCA_030	CNN	D1	-0.0055	0.2409	1.0000	0.6082	1.0000
CNN_PCA_031	CNN	D1	-0.0029	0.5731	1.0000	0.6726	1.0000
CNN_PCA_032	CNN	D1	-0.0066	0.3769	1.0000	0.3336	1.0000
CNN_PCA_033	CNN	D1	-0.0078	0.1458	1.0000	0.1102	1.0000
CNN_PCA_034	CNN	D1	0.0034	0.4357	1.0000	0.8663	1.0000
CNN_PCA_035	CNN	D1	-0.0024	0.6782	1.0000	0.9168	1.0000
CNN_PCA_036	CNN	D1	0.0040	0.5162	1.0000	0.8084	1.0000
CNN_PCA_037	CNN	D1	-0.0015	0.8003	1.0000	0.7822	1.0000
CNN_PCA_038	CNN	D1	-0.0039	0.2559	1.0000	0.6502	1.0000
CNN_PCA_039	CNN	D1	0.0017	0.7490	1.0000	0.5931	1.0000
CNN_PCA_040	CNN	D1	-0.0045	0.2852	1.0000	0.3134	1.0000
CNN_PCA_041	CNN	D1	0.0012	0.8543	1.0000	0.5236	1.0000
CNN_PCA_042	CNN	D1	0.0088	0.1727	1.0000	0.1195	1.0000
CNN_PCA_043	CNN	D1	0.0031	0.6686	1.0000	0.8780	1.0000
CNN_PCA_044	CNN	D1	0.0036	0.3533	1.0000	0.3733	1.0000
CNN_PCA_045	CNN	D1	0.0014	0.8195	1.0000	0.4863	1.0000
CNN_PCA_046	CNN	D1	0.0052	0.4460	1.0000	0.5620	1.0000
CNN_PCA_047	CNN	D1	0.0040	0.4333	1.0000	0.5548	1.0000
CNN_PCA_048	CNN	D1	0.0005	0.9286	1.0000	0.5587	1.0000
CNN_PCA_049	CNN	D1	-0.0017	0.7630	1.0000	0.9452	1.0000
CNN_PCA_050	CNN	D1	-0.0014	0.8384	1.0000	0.8188	1.0000
CNN_PCA_051	CNN	D1	0.0083	0.0149	0.9972	0.0184	0.9953
CNN_PCA_052	CNN	D1	0.0012	0.8377	1.0000	0.9656	1.0000
CNN_PCA_053	CNN	D1	-0.0006	0.8966	1.0000	0.9483	1.0000
CNN_PCA_054	CNN	D1	0.0001	0.9910	1.0000	0.9039	1.0000
CNN_PCA_055	CNN	D1	0.0035	0.4135	1.0000	0.3608	1.0000
CNN_PCA_056	CNN	D1	0.0072	0.1583	1.0000	0.1303	1.0000
CNN_PCA_057	CNN	D1	0.0021	0.6806	1.0000	0.9163	1.0000
CNN_PCA_058	CNN	D1	-0.0036	0.4551	1.0000	0.3453	1.0000
CNN_PCA_059	CNN	D1	0.0014	0.7580	1.0000	0.6459	1.0000
CNN_PCA_060	CNN	D1	0.0045	0.4806	1.0000	0.4206	1.0000
CNN_PCA_061	CNN	D1	-0.0056	0.3986	1.0000	0.3594	1.0000
CNN_PCA_062	CNN	D1	-0.0033	0.4995	1.0000	0.3628	1.0000

continued on next page

Table 17, continued from previous page.

Feature	Family	Tax.	With rarity control			Without rarity control	
			Coef.	p_{boot}	p_{RW}	p_{boot}	p_{RW}
CNN_PCA_063	CNN	D1	0.0003	0.9571	1.0000	0.7450	1.0000
CNN_PCA_064	CNN	D1	0.0022	0.6556	1.0000	0.9067	1.0000
CNN_PCA_065	CNN	D1	0.0034	0.5474	1.0000	0.4557	1.0000
CNN_PCA_066	CNN	D1	-0.0009	0.8465	1.0000	0.6108	1.0000
CNN_PCA_067	CNN	D1	-0.0029	0.4810	1.0000	0.3826	1.0000
CNN_PCA_068	CNN	D1	0.0032	0.2719	1.0000	0.5383	1.0000
CNN_PCA_069	CNN	D1	-0.0109	0.0897	1.0000	0.1015	1.0000
CNN_PCA_070	CNN	D1	-0.0037	0.3473	1.0000	0.2502	1.0000
CNN_PCA_071	CNN	D1	-0.0083	0.1534	1.0000	0.1767	1.0000
CNN_PCA_072	CNN	D1	0.0044	0.2348	1.0000	0.3193	1.0000
CNN_PCA_073	CNN	D1	-0.0074	0.0659	1.0000	0.0654	1.0000
CNN_PCA_074	CNN	D1	-0.0016	0.7215	1.0000	0.8509	1.0000
CNN_PCA_075	CNN	D1	0.0061	0.2432	1.0000	0.1690	1.0000
CNN_PCA_076	CNN	D1	-0.0019	0.7509	1.0000	0.8660	1.0000
CNN_PCA_077	CNN	D1	-0.0018	0.7709	1.0000	0.7275	1.0000
CNN_PCA_078	CNN	D1	-0.0107	0.0731	1.0000	0.0841	1.0000
CNN_PCA_079	CNN	D1	0.0009	0.7697	1.0000	0.4468	1.0000
CNN_PCA_080	CNN	D1	0.0078	0.0451	1.0000	0.0889	1.0000
CNN_PCA_081	CNN	D1	0.0016	0.8207	1.0000	0.7782	1.0000
CNN_PCA_082	CNN	D1	-0.0001	0.9897	1.0000	0.8192	1.0000
CNN_PCA_083	CNN	D1	0.0001	0.9789	1.0000	0.7648	1.0000
CNN_PCA_084	CNN	D1	0.0043	0.2320	1.0000	0.1852	1.0000
CNN_PCA_085	CNN	D1	-0.0058	0.2919	1.0000	0.2257	1.0000
CNN_PCA_086	CNN	D1	-0.0010	0.8503	1.0000	0.6972	1.0000
CNN_PCA_087	CNN	D1	-0.0004	0.9553	1.0000	0.9991	1.0000
CNN_PCA_088	CNN	D1	0.0032	0.5336	1.0000	0.6516	1.0000
CNN_PCA_089	CNN	D1	0.0031	0.6029	1.0000	0.3873	1.0000
CNN_PCA_090	CNN	D1	-0.0006	0.9420	1.0000	0.9416	1.0000
CNN_PCA_091	CNN	D1	0.0039	0.3171	1.0000	0.4223	1.0000
CNN_PCA_092	CNN	D1	0.0017	0.6822	1.0000	0.5090	1.0000
CNN_PCA_093	CNN	D1	0.0042	0.3966	1.0000	0.3747	1.0000
CNN_PCA_094	CNN	D1	0.0031	0.7123	1.0000	0.7155	1.0000
CNN_PCA_095	CNN	D1	-0.0047	0.3422	1.0000	0.2545	1.0000
CNN_PCA_096	CNN	D1	-0.0060	0.0730	1.0000	0.0404	0.9998
CNN_PCA_097	CNN	D1	-0.0003	0.9526	1.0000	0.8423	1.0000
CNN_PCA_098	CNN	D1	-0.0034	0.4138	1.0000	0.4766	1.0000
CNN_PCA_099	CNN	D1	0.0076	0.1529	1.0000	0.2467	1.0000
CNN_PCA_100	CNN	D1	0.0049	0.2841	1.0000	0.2749	1.0000
STYLE_GRAM_PCA_001	STYLE_GRAM	D2	-0.0006	0.9241	1.0000	0.6547	1.0000
STYLE_GRAM_PCA_002	STYLE_GRAM	D2	-0.0127	0.2122	1.0000	0.4023	1.0000
STYLE_GRAM_PCA_003	STYLE_GRAM	D2	0.0089	0.1498	1.0000	0.1150	1.0000
STYLE_GRAM_PCA_004	STYLE_GRAM	D2	0.0018	0.7677	1.0000	0.4527	1.0000
STYLE_GRAM_PCA_005	STYLE_GRAM	D2	-0.0021	0.8112	1.0000	0.8112	1.0000
STYLE_GRAM_PCA_006	STYLE_GRAM	D2	0.0161	0.0248	0.9914	0.0127	0.7428
STYLE_GRAM_PCA_007	STYLE_GRAM	D2	0.0007	0.9203	1.0000	0.9133	1.0000
STYLE_GRAM_PCA_008	STYLE_GRAM	D2	-0.0027	0.6394	1.0000	0.7441	1.0000
STYLE_GRAM_PCA_009	STYLE_GRAM	D2	-0.0090	0.1641	1.0000	0.2860	1.0000
STYLE_GRAM_PCA_010	STYLE_GRAM	D2	0.0003	0.9701	1.0000	0.6531	1.0000
STYLE_GRAM_PCA_011	STYLE_GRAM	D2	-0.0004	0.9428	1.0000	0.9635	1.0000
STYLE_GRAM_PCA_012	STYLE_GRAM	D2	-0.0030	0.6272	1.0000	0.7148	1.0000
STYLE_GRAM_PCA_013	STYLE_GRAM	D2	0.0045	0.5411	1.0000	0.4559	1.0000
STYLE_GRAM_PCA_014	STYLE_GRAM	D2	0.0017	0.7460	1.0000	0.9361	1.0000
STYLE_GRAM_PCA_015	STYLE_GRAM	D2	-0.0028	0.6354	1.0000	0.9513	1.0000

continued on next page

Table 17, continued from previous page.

Feature	Family	Tax.	With rarity control			Without rarity control	
			Coef.	p_{boot}	p_{RW}	p_{boot}	p_{RW}
STYLE_GRAM_PCA_016	STYLE_GRAM	D2	-0.0092	0.2059	1.0000	0.2382	1.0000
STYLE_GRAM_PCA_017	STYLE_GRAM	D2	0.0090	0.2450	1.0000	0.1920	1.0000
STYLE_GRAM_PCA_018	STYLE_GRAM	D2	-0.0004	0.9490	1.0000	0.6672	1.0000
STYLE_GRAM_PCA_019	STYLE_GRAM	D2	0.0082	0.1620	1.0000	0.1040	1.0000
STYLE_GRAM_PCA_020	STYLE_GRAM	D2	-0.0061	0.3472	1.0000	0.2012	1.0000
STYLE_GRAM_PCA_021	STYLE_GRAM	D2	-0.0016	0.7974	1.0000	0.6518	1.0000
STYLE_GRAM_PCA_022	STYLE_GRAM	D2	-0.0076	0.2883	1.0000	0.2079	1.0000
STYLE_GRAM_PCA_023	STYLE_GRAM	D2	0.0086	0.2053	1.0000	0.3090	1.0000
STYLE_GRAM_PCA_024	STYLE_GRAM	D2	-0.0043	0.4759	1.0000	0.6432	1.0000
STYLE_GRAM_PCA_025	STYLE_GRAM	D2	-0.0008	0.9462	1.0000	0.8657	1.0000
STYLE_GRAM_PCA_026	STYLE_GRAM	D2	-0.0009	0.8525	1.0000	0.6771	1.0000
STYLE_GRAM_PCA_027	STYLE_GRAM	D2	0.0016	0.7340	1.0000	0.9719	1.0000
STYLE_GRAM_PCA_028	STYLE_GRAM	D2	-0.0028	0.5246	1.0000	0.5346	1.0000
STYLE_GRAM_PCA_029	STYLE_GRAM	D2	0.0022	0.7393	1.0000	0.7678	1.0000
STYLE_GRAM_PCA_030	STYLE_GRAM	D2	-0.0058	0.2123	1.0000	0.6852	1.0000
STYLE_GRAM_PCA_031	STYLE_GRAM	D2	-0.0014	0.7956	1.0000	0.6547	1.0000
STYLE_GRAM_PCA_032	STYLE_GRAM	D2	-0.0025	0.6259	1.0000	0.6368	1.0000
STYLE_GRAM_PCA_033	STYLE_GRAM	D2	-0.0043	0.3382	1.0000	0.5235	1.0000
STYLE_GRAM_PCA_034	STYLE_GRAM	D2	0.0067	0.3001	1.0000	0.4005	1.0000
STYLE_GRAM_PCA_035	STYLE_GRAM	D2	-0.0131	0.0120	0.9816	0.0071	0.9182
STYLE_GRAM_PCA_036	STYLE_GRAM	D2	-0.0023	0.6675	1.0000	0.4854	1.0000
STYLE_GRAM_PCA_037	STYLE_GRAM	D2	-0.0077	0.1102	1.0000	0.1114	1.0000
STYLE_GRAM_PCA_038	STYLE_GRAM	D2	-0.0018	0.6540	1.0000	0.2258	1.0000
STYLE_GRAM_PCA_039	STYLE_GRAM	D2	0.0073	0.1525	1.0000	0.2040	1.0000
STYLE_GRAM_PCA_040	STYLE_GRAM	D2	0.0009	0.8993	1.0000	0.8450	1.0000
STYLE_GRAM_PCA_041	STYLE_GRAM	D2	0.0036	0.4923	1.0000	0.2398	1.0000
STYLE_GRAM_PCA_042	STYLE_GRAM	D2	-0.0079	0.1888	1.0000	0.2756	1.0000
STYLE_GRAM_PCA_043	STYLE_GRAM	D2	0.0026	0.4819	1.0000	0.5605	1.0000
STYLE_GRAM_PCA_044	STYLE_GRAM	D2	0.0001	0.9760	1.0000	0.6386	1.0000
STYLE_GRAM_PCA_045	STYLE_GRAM	D2	-0.0005	0.9240	1.0000	0.9387	1.0000
STYLE_GRAM_PCA_046	STYLE_GRAM	D2	0.0002	0.9679	1.0000	0.7385	1.0000
STYLE_GRAM_PCA_047	STYLE_GRAM	D2	0.0072	0.1219	1.0000	0.4217	1.0000
STYLE_GRAM_PCA_048	STYLE_GRAM	D2	-0.0163	0.0008	0.3461	0.0004	0.3455
STYLE_GRAM_PCA_049	STYLE_GRAM	D2	0.0043	0.3923	1.0000	0.2259	1.0000
STYLE_GRAM_PCA_050	STYLE_GRAM	D2	0.0005	0.8987	1.0000	0.4318	1.0000