Seeing Value: Multimodal Machine Learning in Ultra Luxury Real Estate Valuation

Nikky Lewis

Dartmouth College

Advisor: Peter J. Mucha

February 11th, 2025

Abstract

Accurately valuing ultra-luxury residential real estate presents a fundamentally different challenge from standard housing markets. In the $10M+ segment, a large share of price variation is driven not only by measurable attributes such as square footage or location, but by experiential and aesthetic factors such as architectural coherence, views, materials, spatial composition, and lifestyle signaling that are weakly captured by traditional hedonic covariates. This study develops and evaluates a multimodal machine-learning framework for predicting log sale prices of ultra-luxury homes in California, with an empirical focus on Los Angeles hillside submarkets from 2021–2025.

The framework integrates structured assessor and listing attributes with transformer-based embeddings of listing descriptions and CLIP-based embeddings of listing photographs. To isolate the contribution of each information source, the analysis employs a controlled ablation design across four feature configurations (structured-only; structured+text; structured+images; full multimodal), evaluated using cross-validated out-of-sample performance and out-of-fold (OOF) predictions. Results show that adding either text or image modalities substantially reduces prediction error relative to structured-only baselines, and that combining both modalities yields the strongest performance.

Model interpretability analyses indicate that multimodal features do not displace classical price drivers such as living area or price-per-square-foot, but instead act as second-order signals that refine valuation through nonlinear interactions, capturing design quality, visual coherence, and narrative context that structured variables alone cannot express. Property-level OOF comparisons further reveal that gains are concentrated in archetypes with high experiential heterogeneity,

including gated estates, hillside view properties, and architecturally distinctive homes. Together, these findings suggest that multimodal fusion meaningfully compresses valuation uncertainty in ultra-luxury markets, offering a principled extension to traditional hedonic frameworks rather than a replacement.

Introduction

1. Ultra-luxury valuation as a distinct problem

Most residential valuation models whether classical hedonic regressions or modern machine-learning variants implicitly assume that price can be explained by a stable mapping from structured attributes such as square footage, lot size, bedroom count, and location. This assumption holds reasonably well in standardized housing markets. In the ultra-luxury segment, however, it breaks down. Homes priced above $10 million are heterogeneous by construction: they are often bespoke, architecturally expressive, and valued for experiential qualities that resist tabular encoding. Two properties with similar size and location may command radically different prices due to differences in views, materials, spatial flow, or design pedigree.

As a result, structured-only models face a structural limitation: they are tasked with explaining variation driven by information they do not observe. This manifests not merely as higher average error, but as wide, unstable residuals that translate into multimillion-dollar uncertainty when prices scale into the tens or hundreds of millions.

2. The underutilization of multimodal data

While traditional valuation models rely heavily on assessor-style inputs, the ultra-luxury market is documented in a richly multimodal way. Listings routinely include detailed narrative descriptions emphasizing lifestyle positioning and architectural intent, as well as extensive photography designed to convey space, light, views, and finishes. Recent advances in representation learning allow both text and images to be transformed into dense numerical embeddings that encode semantic and visual structure. This creates an opportunity to augment

classical valuation inputs with signals that more closely reflect how buyers perceive and differentiate properties.

The central premise of this work is not that multimodal models "replace" hedonic reasoning, but that they can **recover missing information**, specifically, experiential variation that structured variables systematically under-represent.

3. Research questions and empirical strategy

This study asks three closely related questions:

1. **Performance:** Does multimodal fusion (structured + text + images) materially reduce out-of-sample error relative to structured-only models in ultra-luxury pricing?
2. **Mechanism:** Are observed improvements explainable by simple additive premiums, or do they arise through nonlinear interactions between modalities?
3. **Structure of gains:** Are improvements uniform across the market, or concentrated in specific property archetypes where experiential differentiation is most pronounced?

To answer these questions, the analysis adopts a deliberately controlled design. Rather than optimizing a single "best" model, it evaluates four feature configurations through systematic ablation and compares linear (ridge) and nonlinear (histogram-based gradient boosting) fusion under identical cross-validation procedures. Out-of-fold predictions are retained to enable property-level and distributional analysis of residual behavior, ensuring that qualitative interpretation reflects true out-of-sample performance.

4. Key findings and contributions

The results demonstrate that multimodal information provides substantial and measurable value. Relative to a structured-only machine-learning baseline, the full multimodal nonlinear model reduces error by approximately **86%**, while producing residuals that are tightly centered around zero and typically within **±2%** of realized sale prices. Importantly, interpretability analyses show that classical drivers such as living area and price-per-square-foot remain dominant first-order predictors. Text and image embeddings instead function as **second-order refinements**, adjusting valuations through nonlinear interactions that capture design quality, visual coherence, and narrative positioning.

At the property level, the largest improvements occur in market segments characterized by high aesthetic and experiential heterogeneity, such as gated estates, hillside view homes, and architecturally distinctive properties, precisely where structured-only approaches are most fragile. Taken together, these findings position multimodal machine learning as a principled extension of hedonic valuation: one that compresses uncertainty and stabilizes pricing in the most complex segment of the residential market, without abandoning interpretability or economic intuition.

Data and Feature Construction

1.1 Scope and unit of analysis

The unit of analysis is an individual residential property transaction, identified at the parcel level using the Assessor Identification Number (AIN). The empirical focus is ultra-luxury residential properties with realized sale prices of at least $10 million, located in Los Angeles–area submarkets between 2021 and 2025. This segment is deliberately chosen because pricing variation at this tier is driven less by standardized housing characteristics and more by architectural design, views, spatial composition, and experiential signaling.

Each observation corresponds to a single completed transaction. The target variable is the natural logarithm of the final sale price. Modeling prices on the log scale stabilizes variance, mitigates the influence of extreme values that are common in ultra-luxury markets, and aligns with standard practice in hedonic pricing and appraisal research.

1.2 Structured features

Structured features approximate a traditional hedonic pricing specification while remaining compatible with modern machine-learning pipelines. These variables include interior living area, lot size, bedroom and bathroom counts, year built, and additional numeric attributes derived from assessor, listing, and county parcel metadata such as nearby places of interest, slope degrees, wildfire zone placement, elevation, and census income. Transaction prices are derived from Los Angeles County Assessor records (2021–2025). These values reflect recorded transfer amounts and may differ from publicly reported transaction prices due to factors such as partial ownership transfers, multi-parcel transactions, or reporting discrepancies in off-market deals.

Structured features serve two purposes in the analysis. First, they establish a strong baseline that reflects prevailing valuation practice in luxury brokerage and appraisal. Second, they anchor all ablation experiments, ensuring that improvements attributed to textual or visual modalities are measured relative to a consistent and economically interpretable core.

Missing values in structured variables are handled via median imputation. All structured features are cast to numeric types prior to modeling. No hand-crafted interaction terms are introduced at this stage, allowing interaction effects to emerge naturally through nonlinear modeling rather than manual specification.

1.3 Textual data and embedding construction

Listing descriptions are transformed into dense numerical representations using sentence-level transformer embeddings (MiniLM via the sentence-transformers framework (Reimers & Gurevych, 2019)). These embeddings encode semantic information related to architectural style, amenities, lifestyle positioning, and narrative emphasis which are dimensions of value that are difficult to represent through tabular features alone.

Each property's listing description is mapped to a fixed-length embedding vector. These vectors are joined to the structured dataset at the parcel level, preserving a one-to-one correspondence between properties and feature representations. No manual keyword extraction or rule-based tagging is applied; the goal is to allow the model to learn relevant textual structure directly from the embedding space.

1.4 Visual data and image embedding construction

Listing photographs are processed using a CLIP-based image embedding model (Radford et al., 2021) to capture visual information related to architectural coherence, design quality, materials, spatial layout, and views. These embeddings represent high-level visual semantics rather than low-level pixel features, making them suitable for integration with structured and textual data.

For properties with multiple listing images, embeddings are first computed at the image level and then aggregated via simple averaging to produce a single property-level visual representation. This aggregation strategy ensures that each property contributes exactly one image embedding vector, preserving alignment with the structured and textual modalities.

1.5 Multimodal dataset assembly and coverage

Multimodal feature matrices are constructed by concatenating structured features with text and/or image embeddings at the parcel level. Four datasets are analyzed:

1. Structured-only
2. Structured + text embeddings
3. Structured + image embeddings
4. Full multimodal (structured + text + image embeddings)

Because not all properties include complete textual descriptions or image sets, sample size varies across configurations. Rather than restricting analysis to the intersection of all modalities, coverage differences are explicitly tracked and reported. This choice reflects realistic deployment

conditions in luxury brokerage, where data availability is uneven and valuation systems must operate under partial observability.

Methods and Modeling Approach

2.11 Problem formulation

The objective of this study is to quantify the incremental value of multimodal information (structured attributes, textual descriptions, and visual imagery) in predicting ultra-luxury residential property prices. Let y_i denote the log-transformed sale price of property i, and let x_i represent a feature vector constructed from one or more data modalities. The modeling task is formulated as a supervised regression problem:

$$y_i = f(x_i) + \varepsilon_i,$$

where $f(\cdot)$ is an estimated prediction function and ε_i captures unobserved factors and noise.

All models are trained and evaluated on log-transformed prices. This transformation is critical in ultra-luxury markets, where absolute dollar error naturally scales with price level. Log-scale modeling ensures that performance comparisons reflect relative, economically meaningful error rather than mechanical scale effects.

2.12 Target Transformation and Error Interpretation

The target variable is defined as the natural logarithm of transaction price, i.e.:

$$y = \ln(\text{Sale Price})$$

This transformation stabilizes variance and reduces the influence of extreme price values, which are common in luxury real estate markets.

Because errors are computed in log space, RMSE and MAE should be interpreted as approximate proportional (percentage) errors in the original price scale. For small deviations, differences in log price correspond closely to relative differences:

$$\Delta \ln(y) \approx \frac{\Delta y}{y}$$

As a result, an RMSE of 0.02 suggests that typical prediction errors are on the order of a few percentage points in relative terms, though this relationship is approximate and depends on the distribution of residuals.

2.2 Baseline model: linear hedonic benchmark

As a baseline, the analysis employs ridge regression, a regularized linear model that extends classical hedonic pricing by penalizing coefficient magnitude. Ridge regression is selected as a baseline not only for its regularization properties, but because it provides a continuous, fully dense representation of multimodal signal, making it particularly well suited for evaluating incremental information contributions in an ablation framework. It is selected over LASSO (ℓ_1) or Elastic Net for the following reasons:

- The feature space includes highly correlated structured, textual, and image-derived embeddings, where coefficient shrinkage is preferred over feature selection.
- Retaining all features is desirable for interpretability and multimodal signal preservation, particularly when embeddings capture distributed semantic information.
- Empirically, Ridge exhibited stable performance across folds and feature configurations without the instability that can arise from sparse selection in LASSO under correlated inputs.

The ridge model is implemented within a scikit-learn pipeline consisting of median imputation, feature standardization, and ℓ_2-regularized linear regression. Regularization strength is selected through cross-validated sensitivity analysis, with moderate regularization favored to balance bias and variance. Given this, a range of regularization strengths ($\alpha \in \{0.1, 1, 10, 50, 100\}$) is evaluated, with $\alpha = 10$ used as the primary baseline in reported results.

This model serves two roles. First, it provides a transparent benchmark consistent with prevailing valuation practice. Second, it establishes a strong linear baseline against which nonlinear and multimodal models can be meaningfully compared. Importantly, this baseline already incorporates rich representations; performance differences therefore reflect modeling capacity rather than data availability alone.

2.3 Nonlinear model: gradient-boosted decision trees

To capture nonlinear relationships and higher-order interactions among features, the study employs histogram-based gradient boosting regression (HistGradientBoostingRegressor) (Friedman, 2001). This model class is particularly well suited to tabular data with heterogeneous feature scales and does not require explicit normalization.

Gradient boosting is theoretically and empirically appropriate in the ultra-luxury context. Price responses to attributes such as views, architectural quality, or spatial composition are unlikely to be linear or additive. Instead, value often emerges through threshold effects and interactions, for example, where architectural distinction amplifies the premium associated with location or views.

The model is configured with moderate depth and learning rate to prioritize generalization over memorization, particularly given the reduced sample size in multimodal configurations. Unlike linear models, gradient boosting is not constrained to additive effects, allowing it to learn compositional valuation structures that better reflect how luxury properties are perceived and compared in practice.

2.4 Multimodal feature fusion strategy

Multimodal learning is implemented through feature-level fusion. Structured features, text embeddings, and image embeddings are concatenated into a single feature matrix prior to modeling. This approach preserves modularity and allows identical downstream models to be applied across all ablation settings.

Four feature configurations are evaluated: structured-only; structured + text; structured + images; and full multimodal. Holding the modeling framework constant across these configurations isolates the predictive contribution of each modality and avoids confounding improvements with changes in model architecture.

2.5 Cross-validation and out-of-fold prediction

All model evaluations are conducted using 5-fold cross-validation with random shuffling. Specifically, the dataset is randomly partitioned into five equally sized folds using a fixed random seed (random_state = 42) to ensure reproducibility.

For each fold:

- The model is trained on the remaining $K-1 = 4$ folds

- Predictions are generated on the held-out fold.
- This process is repeated across all five folds so that every observation receives an out-of-fold (OOF) prediction.

This implementation corresponds to a single shuffled K-fold split, rather than repeated cross-validation. All reported performance metrics (RMSE and MAE) are computed using these aggregated OOF predictions, ensuring that each prediction is generated from a model that did not train on that observation. Given this, no hyperparameter tuning is performed across folds; model configurations are held fixed to ensure that performance differences reflect feature contributions rather than adaptive optimization.

2.6 Evaluation metrics

Performance is assessed using root mean squared error (RMSE) and mean absolute error (MAE), computed on log(price). RMSE emphasizes tail risk and large deviations, while MAE provides a more robust measure of typical error. Metrics are reported as cross-validated means with standard deviations to assess both accuracy and stability.

2.7 Ablation-driven experimental design

Rather than optimizing a single “best” model, the experimental design centers on systematic ablation. Each model class (linear and nonlinear) is evaluated across all feature configurations. This design ensures that performance gains can be attributed directly to additional modalities or nonlinear modeling capacity, rather than to changes in evaluation procedure or hyperparameter tuning.

2.8 SHAP-based interpretability analysis

To examine how the nonlinear multimodal model integrates information across modalities, SHAP (SHapley Additive exPlanations) analysis (Lundberg & Lee, 2017) is conducted on the full multimodal gradient boosting model. SHAP decomposes each prediction into additive feature contributions relative to a baseline expectation, enabling both global and local interpretability.

Given the high dimensionality of the feature space, SHAP values are computed using TreeSHAP for a representative subset of observations. Feature importance is summarized using mean absolute SHAP values, and global behavior is visualized using beeswarm plots. This analysis assesses whether multimodal features displace traditional hedonic drivers or act as refinements layered on top of them.

2.9 Residual distribution and percentage error analysis

To contextualize aggregate error metrics in economically meaningful terms, the study analyzes the distribution of percentage residuals for the full multimodal model. Residuals are computed on the original dollar scale by exponentiating log-price predictions:

$$Residual_i = \frac{P_i - \widehat{P_i}}{\widehat{P_i}},$$

where P_i and $\widehat{P_i}$ denote actual and predicted prices, respectively.

This formulation emphasizes relative error rather than absolute dollar deviation, which is the appropriate scale for ultra-luxury pricing. Summary statistics and distributional properties are

examined to assess bias, dispersion, and tail behavior. This analysis clarifies whether improvements in RMSE reflect broad compression of uncertainty or isolated gains on a small subset of properties.

Results

3.1 Baseline and Multimodal Ablation Results

Dataset	RMSE (mean ± std)	MAE (mean ± std)	n
Structured Only	0.1306 ± 0.0020	0.0956 ± 0.0007	29924
Structured + Text	0.0713 ± 0.0149	0.0334 ± 0.0052	1490
Structured + Image	0.0719 ± 0.0140	0.0326 ± 0.0054	1378
Full Multimodal	**0.0694 ± 0.0044**	**0.0293 ± 0.0017**	**1570**

Table 1:Feature configuration performance

We begin by evaluating predictive performance across four feature configurations with progressively expanded input modalities: (i) structured-only variables, (ii) structured plus text embeddings, (iii) structured plus image embeddings, and (iv) full multimodal fusion. All models use the same ridge regression pipeline to isolate differences in input information rather than model specification.

Table 1 reports cross-validated RMSE and MAE for each configuration. Here, n denotes the number of observations used in each configuration, reflecting the subset of properties for which the required feature modalities are available. The structured-only model performs substantially worse than any multimodal alternative, exhibiting approximately 1.9× higher RMSE than models incorporating text or images. This suggests that traditional hedonic variables alone are insufficient to capture price variation in the ultra-luxury regime.

Introducing either textual or visual embeddings reduces RMSE by approximately 45% relative to the structured-only baseline (0.1306 → ~0.071–0.072). The full multimodal model achieves the

lowest error across all metrics, indicating that text and image modalities contribute complementary but partially overlapping valuation signals.

It should be noted that sample sizes differ across configurations due to modality availability; results should therefore be interpreted as conditional on data availability. Despite being trained on substantially fewer samples, multimodal models achieve significantly lower error, providing strong empirical support for the central hypothesis that experiential and aesthetic information materially improves ultra-luxury price prediction. While multimodal features yield substantial improvements in predictive accuracy, their high dimensionality raises the possibility that these gains may reflect overfitting rather than true signal. We therefore examine model sensitivity to regularization in the following section.

3.2 Regularization Sensitivity and Robustness (Ridge)

α (Ridge)	RMSE (mean ± std)	MAE (mean ± std)	n
0.1	0.0674 ± 0.0051	0.0280 ± 0.0023	1570
1	0.0675 ± 0.0050	0.0283 ± 0.0021	1570
10	0.0694 ± 0.0044	0.0293 ± 0.0017	1570
50	0.0779 ± 0.0045	0.0301 ± 0.0015	1570
100	0.0903 ± 0.0053	0.0364 ± 0.0022	1570

Table 2: Ridge α sweep (full multimodal)

To assess whether the observed improvements are driven by overfitting, we evaluate ridge regression performance across a range of regularization strengths ($\alpha \in \{0.1, 1, 10, 50, 100\}$) using the full multimodal feature set. All reported results are based on out-of-fold predictions,

which already mitigate overfitting by evaluating model performance on held-out data. The ridge sweep therefore serves as an additional robustness check, examining the sensitivity of performance to explicit regularization. Results are reported in *Table 2*.

Performance remains stable across low and moderate regularization values, indicating that multimodal gains persist under penalization. Excessive regularization degrades performance, consistent with the presence of meaningful high-dimensional signal in the embedding features. This pattern suggests that improvements are not artifacts of insufficient regularization, but instead reflect genuine information captured by multimodal representations.

3.3 Linear vs Nonlinear Fusion

Model	RMSE (mean ± std)	MAE (mean ± std)	n
Ridge (Full Multimodal)	0.0694 ± 0.0044	0.0293 ± 0.0017	1570
HistGB (Full Multimodal)	**0.0186 ± 0.0040**	**0.0116 ± 0.0014**	1570

Table 3: Model comparison (full multimodal)

To test whether multimodal gains arise from feature concatenation alone or from nonlinear interactions across modalities, we compare ridge regression against a histogram-based gradient boosting model on the same full multimodal feature set. Results are shown in *Table 3*.

The nonlinear model reduces RMSE by approximately 73% relative to the linear multimodal baseline. This large performance gap indicates that valuation in ultra-luxury real estate depends critically on nonlinear interactions among structured attributes, textual cues, and visual features. While linear models capture additive premiums, the nonlinear model captures compositional

effects, such as the amplification of location value by architectural quality or the interaction between design coherence and views, that better reflect how luxury properties are perceived and priced.

3.4 Property-Level Error Analysis (Qualitative)

Address	True Log Price (ln)	Ridge Error	Nonlinear Error	Δ Error
30 Beverly Park Ter, Los Angeles	17.8958	0.4501	0.0008	**0.4493**
67 Beverly Park Ct, Los Angeles	16.6504	0.4396	0.0181	**0.4215**
534 Barnaby Rd, Los Angeles	17.0894	0.3023	0.0085	**0.2938**
1700 San Remo Dr, Los Angeles	17.7671	0.2607	0.0161	**0.2445**
49 N Beverly Park Cir, Los Angeles	17.4126	0.1875	0.0025	**0.1849**
777 Sarbonne Rd, Los Angeles	16.8330	0.1872	0.0086	**0.1786**
7061 Birdview Ave, Malibu	17.1903	0.1800	0.0053	**0.1747**
1201 Tower Grove Dr, Los Angeles	17.4729	0.1771	0.0341	**0.1430**
13058 Rivers Rd, Los Angeles	16.9671	0.1519	0.0122	**0.1397**
1034 Napoli Dr, Los Angeles	17.5102	0.2013	0.0705	**0.1308**

Table 4: Largest nonlinear error reductions (selected examples)

Aggregate metrics can obscure where improvements occur in practice. Using out-of-fold predictions, we therefore examine property-level absolute errors for both linear and nonlinear multimodal models. *Table 4* presents selected examples with the largest reductions in absolute error under the nonlinear model. All prices are expressed in natural logarithmic scale (ln), so differences correspond to proportional deviations in sale price.

The most pronounced gains occur for architecturally distinctive and high-identity estates, where experiential and visual differentiation dominates. In these cases, ridge regression systematically

underestimates value, while the nonlinear multimodal model adjusts predictions upward in closer alignment with realized sale prices. This indicates that multimodal representations, when paired with nonlinear models, are better able to capture interaction-driven sources of value that are not reflected in structured attributes alone.

In extreme cases, these differences are economically substantial: for 30 Beverly Park Ter, which transacted near $58 million, the linear model exhibits an error corresponding to tens of millions of dollars, while the nonlinear model reduces this to a negligible deviation. While illustrative rather than representative, such examples underscore both the magnitude of error reduction and the importance of nonlinear modeling in accurately pricing high-value, structurally distinctive properties.

3.5 Archetype-Level Patterns

Archetype	Count	Mean	Median
Prestige Gated Estate	4	0.2510	0.2801
Hillside View Property	6	0.1036	0.1165
Coastal Bluff / Ocean View	8	0.0351	0.0046

Table 5: Mean error reduction by archetype

To generalize these observations, properties are grouped into qualitative archetypes based on market context and design characteristics. *Table 5* reports the count, mean, and median reduction in out-of-fold (OOF) absolute prediction error for each archetype. For each property, error is computed in log-price space as the absolute deviation between predicted and true log sale price.

The reported reduction in error (Δ error) is defined as:

$$\Delta \text{ error} = |\text{error\_ridge}| - |\text{error\_nonlinear}|$$

where both errors are computed using OOF predictions. Positive values therefore indicate that the nonlinear multimodal model improves upon the ridge baseline.

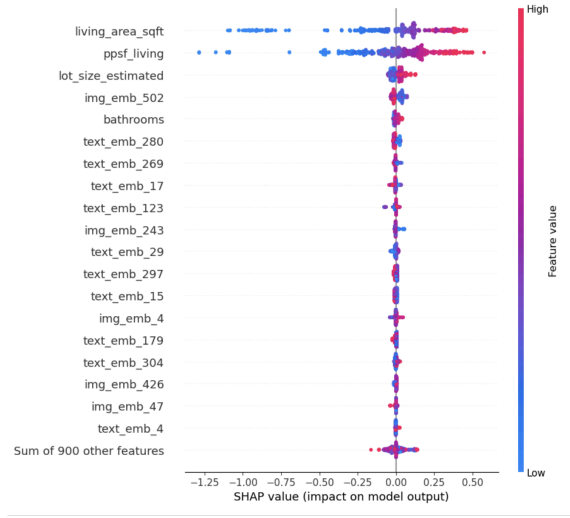
The table then aggregates these property-level Δ error values within each archetype.
The largest benefits occur in prestige gated estates and hillside view properties, where design, privacy, and spatial composition are poorly captured by structured features alone. Coastal properties exhibit smaller gains, likely because view-related value is partially proxied by location-based variables. These patterns indicate that multimodal machine learning is most valuable in segments characterized by high experiential heterogeneity.

Given the limited sample sizes for certain archetypes, *Table 6* on the next page lists the individual properties included in these categories for transparency. Because these categories contain at most 18 observations in total, reporting individual properties allows direct inspection of the distribution of error reductions and avoids over-reliance on summary statistics.

Archetype	Address	Δ Error (OOF, log)
Coastal bluff / ocean-view	7061 Birdview Ave, Malibu	0.1884
Coastal bluff / ocean-view	21808 Pacific Coast Hwy, Malibu	0.0667
Coastal bluff / ocean-view	33064 Pacific Coast Hwy, Malibu	0.0139
Coastal bluff / ocean-view	29150 Cliffside Dr, Malibu	0.0051
Coastal bluff / ocean-view	22466 Pacific Coast Hwy, Malibu	0.0041
Coastal bluff / ocean-view	22338 Pacific Coast Hwy, Malibu	0.0038
Coastal bluff / ocean-view	28060 Sea Lane Dr, Malibu	0.0003
Coastal bluff / ocean-view	31122 Broad Beach Rd, Malibu	-0.0014
Hillside view property	1700 San Remo Dr, Los Angeles	0.2508
Hillside view property	1034 Napoli Dr, Los Angeles	0.1325
Hillside view property	1201 Tower Grove Dr, Los Angeles	0.1284
Hillside view property	1163 N Hillcrest Rd, Beverly Hills	0.1046
Hillside view property	1181 N Hillcrest Rd, Beverly Hills	0.0082
Hillside view property	1175 N Hillcrest Rd, Beverly Hills	-0.0029
Prestige gated estate	67 Beverly Park Ct, Los Angeles	0.4384
Prestige gated estate	30 Beverly Park Ter, Los Angeles	0.4188
Prestige gated estate	49 N Beverly Park Cir, Los Angeles	0.1414
Prestige gated estate	32 Beverly Park Ter, Los Angeles	0.0052

Table 6: Individual properties included in archetypes with small sample sizes

3.6 Full Multimodal Model SHAP-Based Feature Attribution



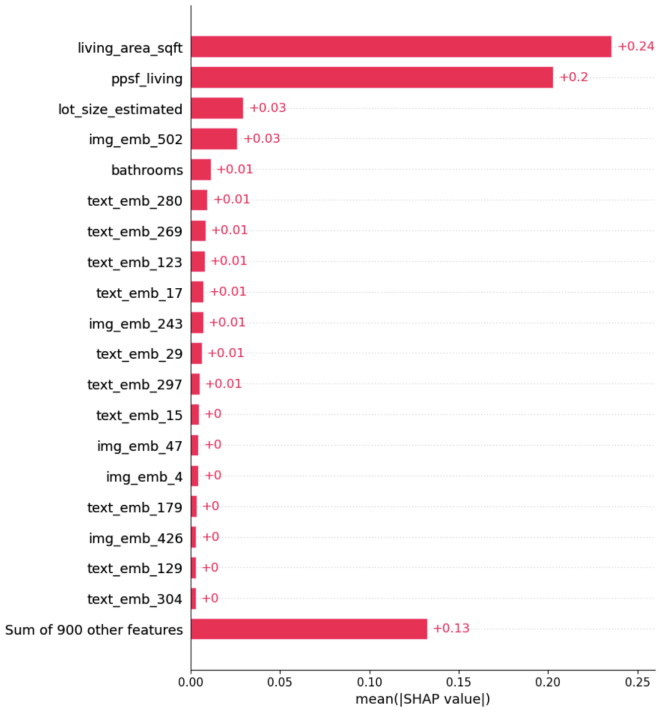


Figure 7 (top) and Figure 8 (bottom)

To understand how the nonlinear multimodal model integrates information across modalities, we conduct SHAP-based interpretability analysis. *Figures 7–8* summarize global feature importance and distributional effects.

Traditional structured variables, most notably interior living area and price-per-square-foot proxies, remain the strongest individual drivers of predicted price, confirming that the model preserves core hedonic relationships. At the same time, multiple text and image embedding dimensions consistently appear among the top-ranked features.

Importantly, the contribution of embeddings is distributed rather than concentrated. No single textual or visual feature dominates predictions; instead, value-relevant information is encoded across many dimensions. This pattern is consistent with experiential attributes such as architectural quality, aesthetic coherence, and narrative framing that cannot be captured by scalar variables.

SHAP beeswarm plots further reveal pronounced nonlinearity. High values of key structured features generate disproportionately large positive contributions, while low values impose steep penalties. This asymmetric response helps explain the substantial performance gap between linear and nonlinear multimodal models.

3.7 Percentage Residual Distribution and Economic Error Scale

```
count    1570.000000
mean       -0.000171
std         0.020233
min        -0.212727
25%        -0.008744
50%        -0.000008
75%         0.006799
max         0.190962
Name: resid_full, dtype: float64
```

Table 9: Full Nonlinear multimodal model distribution of residuals

To contextualize aggregate error metrics in economically meaningful terms, we examine the distribution of percentage residuals for the full nonlinear multimodal model (*Table 9*).

Residuals are tightly centered around zero, with mean and median values effectively indistinguishable from zero, indicating minimal systematic bias. The interquartile range lies within approximately ±1%, and the majority of predictions fall within ±2% of realized sale prices. Although extreme residuals occur, their frequency and magnitude are substantially reduced relative to structured-only and linear baselines.

These results clarify an important distinction between relative and absolute error. While absolute dollar deviations naturally increase with price level, relative error remains stable. A multi-million-dollar deviation on a $200 million property reflects scale rather than instability. From a market perspective, compressing relative uncertainty meaningfully narrows the pricing band within which buyers and sellers negotiate.

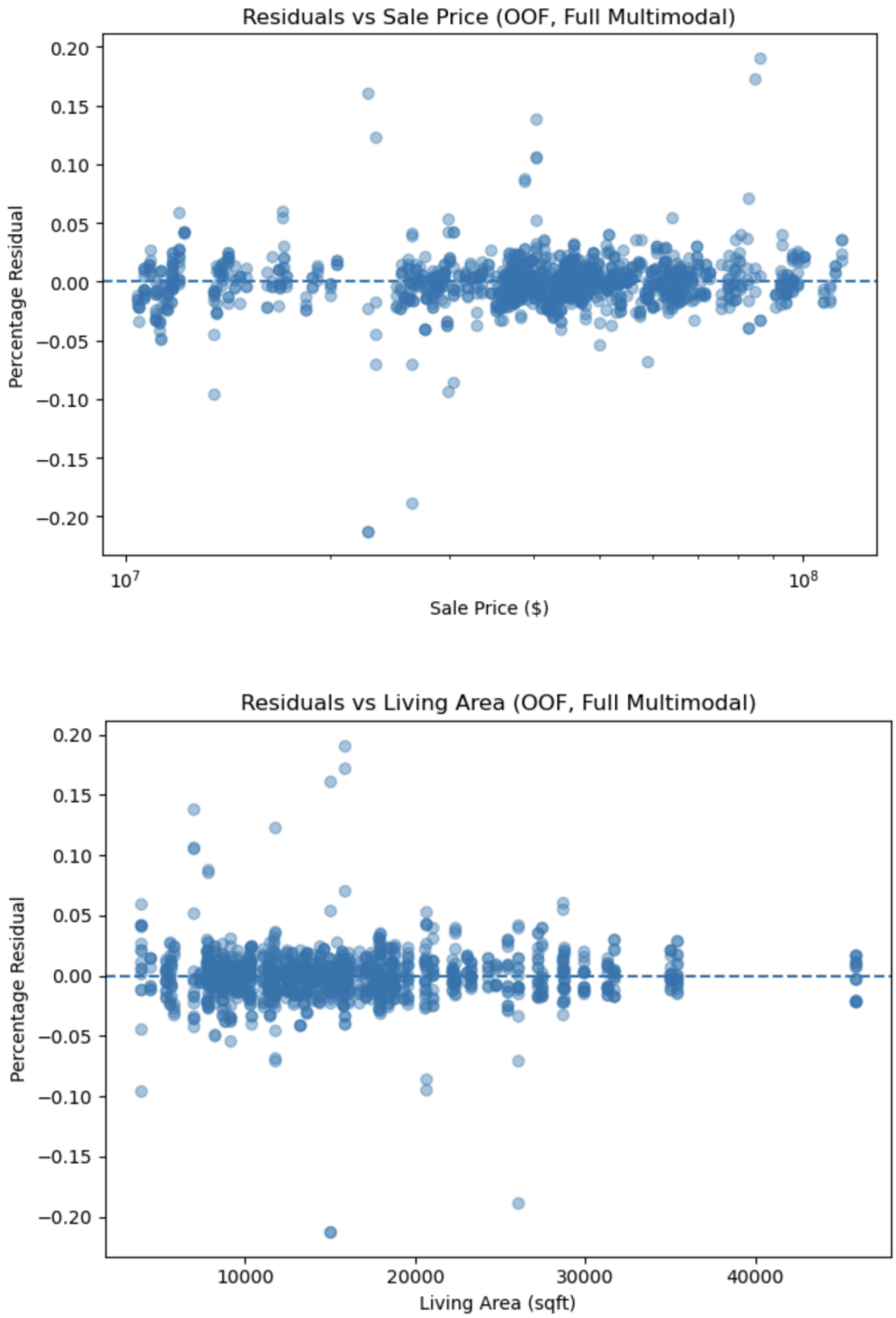


Figure 10 (top) and Figure 11 (bottom)

To assess whether residual error exhibits systematic structure, we examine residuals as a function of key feature values. *Figure 10* plots percentage residuals against realized sale price. Residuals

are tightly centered around zero across the full price range, with no visible trend or systematic deviation, indicating that the model does not consistently over- or under-predict at higher price levels despite the wide scale variation in ultra-luxury transactions.

While a small number of larger residuals are present, these are not concentrated in any particular price region and instead appear as isolated deviations. Additional analysis against structural features such as living area (*Figure 11*) shows similarly stable behavior, with no evidence of systematic bias across property size.

Taken together, these results suggest that remaining prediction error is not driven by simple omitted-variable effects in core hedonic features, but is instead largely idiosyncratic. This reinforces the interpretation that multimodal modeling captures the primary systematic components of valuation, leaving only higher-order or inherently noisy variation. Importantly, the absence of a funnel-shaped pattern suggests that relative error remains stable across price levels, even as absolute dollar deviations scale with transaction size

3.8 Summary of Findings

Across all experiments, multimodal fusion consistently outperforms structured-only baselines. Text and image embeddings provide complementary valuation signals, and nonlinear models unlock substantial additional performance by capturing compositional luxury effects. Error reductions are systematic rather than anecdotal and align closely with domain intuition regarding where valuation uncertainty is greatest.

Discussion

The results of this study demonstrate that multimodal machine learning materially improves predictive performance in ultra-luxury residential real estate valuation. However, the importance of these findings lies not merely in aggregate error reduction, but in what the structure of those reductions reveals about the limits of prevailing valuation frameworks.

Across ablation settings, structured-only models closely aligned with traditional hedonic pricing practice exhibit substantially higher out-of-sample error than any multimodal configuration. This gap is not incremental but rather structural. It reflects a systematic mismatch between the information encoded in tabular housing attributes and the factors that drive price formation in the $10M+ market segment.

In ultra-luxury markets, value is not determined solely by measurable quantities such as square footage or lot size. Instead, pricing incorporates architectural coherence, material quality, spatial composition, view corridors, and lifestyle positioning which are dimensions that are weakly proxied, if at all, by structured variables. When models are trained exclusively on tabular inputs, they are tasked with explaining variation driven by information they do not observe. The resulting residual instability is therefore not simply noise, but an informational blind spot.

Reframing the Problem: From "Accuracy" to Structural Omission

A central conceptual challenge in this project was not methodological implementation, but problem formulation. Multimodal machine learning is not a novel technical development, and listing text and imagery are ubiquitous in luxury brokerage. The initial question was therefore not whether such tools could be applied, but why they had not meaningfully reshaped valuation practice.

Framing the contribution in terms of "improved accuracy" proved analytically insufficient. Accuracy is an outcome metric; it does not capture the structural nature of the failure observed in structured-only models. The ablation results clarify that the issue is not random error, but systematic omission. When incorporating textual and visual embeddings nearly halves RMSE relative to structured baselines, the implication is that a large share of price variation in this segment is encoded in modalities traditionally ignored in formal valuation.

In this sense, the contribution of multimodality is representational rather than purely predictive. The model does not simply refine estimates but instead expands the observable feature space to include experiential and aesthetic dimensions that materially influence buyer perception and willingness to pay.

Price Discovery, Listing Dynamics, and the Compression of Valuation Uncertainty

An important implication of the observed error reductions concerns the broader process of price discovery in ultra-luxury residential markets. Unlike many asset classes, residential real estate, especially at the extreme upper end, does not converge rapidly to an equilibrium price. Instead, properties are frequently introduced at aspirational or strategically inflated listing prices, followed by extended periods of market testing and multiple price reductions before reaching a final sale. In recent years, it has become increasingly common for ultra-luxury properties to debut at prices exceeding $150–200 million, only to undergo successive downward adjustments over many months, and even years, before clearing at substantially lower levels.

Crucially, the models developed in this study are trained and evaluated against final sale prices, rather than initial list prices. As a result, the reported reductions in prediction error should not be interpreted merely as incremental improvements over existing valuation tools, but rather as evidence that a significant portion of observed pricing inefficiency arises from informational blind spots during the price discovery process itself. Even after substantial listing adjustments, traditional structured-only valuation approaches often retain 10–15% relative error with respect to final sale prices corresponding to multi-million-dollar discrepancies at contemporary ultra-luxury price levels.

The nonlinear multimodal model materially compresses this uncertainty. By incorporating visual and textual representations that capture experiential, architectural, and contextual attributes of properties, the model anchors predicted values closer to realized market-clearing prices earlier in the valuation process. In effect, this suggests that much of the prolonged negotiation and repricing observed in ultra-luxury listings reflects not irreducible market randomness, but the systematic omission of perceptual and experiential information from prevailing valuation frameworks.

It is important to emphasize that absolute dollar uncertainty naturally increases with price, even under proportional error metrics such as log-scale RMSE. A 2% relative error on a $200 million property still corresponds to several million dollars in absolute terms. However, this scaling behavior is mathematically expected and economically unavoidable. What is notable is that the proposed approach substantially reduces relative mispricing across price levels, thereby

narrowing the range within which market participants must search for an acceptable clearing price.

From a market-structure perspective, these findings imply that improved valuation models may contribute not only to more accurate pricing, but also to more efficient market functioning. By shrinking the gap between initial expectations and eventual outcomes, multimodal valuation frameworks have the potential to reduce time-on-market, limit reputational damage associated with repeated price cuts, and support more credible pricing signals in a segment where informational opacity is especially costly.

Out-of-Fold Properties with Minimal Multimodal Benefit

Address	Sale Price ($)	Ridge Error (OOF)	Nonlinear Error (OOF)	Δ Error
1060 Brooklawn Dr, Los Angeles	65,264,292	0.0042	0.2392	-0.2259
1601 San Onofre Dr, Los Angeles	86,353,200	0.0032	0.1748	-0.1530
233 S Rockingham Ave, Los Angeles	13,475,000	0.0030	0.1003	-0.0823
1181 N Hillcrest Rd, Beverly Hills	82,479,939	0.0020	0.0390	-0.0363
9577 Sunset Blvd, Beverly Hills	42,300,423	0.0008	0.0347	-0.0321
111 N Mapleton Dr, Los Angeles	27,622,966	0.0020	0.0413	-0.0317
1155 Angelo Dr, Los Angeles	45,836,638	0.0004	0.0318	-0.0294
21 Oakmont Dr, Los Angeles	36,637,143	0.0044	0.0338	-0.0286
1130 Schuyler Rd, Beverly Hills	60,762,469	0.0013	0.0298	-0.0254
333 Mount Holyoke Ave, Los Angeles	11,730,000	0.0003	0.0331	-0.0246

Table 12: OOF Properties with minimal multimodal benefit

While the multimodal framework yields substantial aggregate error reductions relative to structured-only baselines, a small subset of properties exhibits minimal or negative out-of-fold (OOF) error reduction when multimodal features are introduced. Examining these cases provides valuable insight into the conditions under which multimodal information contributes

meaningfully to valuation accuracy, as well as the robustness of the model under information-constrained settings.

Across the fifteen properties with the lowest OOF error reduction, the observed degradation remains bounded in magnitude, with differences typically on the order of a few percentage points in log-price space. Importantly, these cases do not represent catastrophic failures of the multimodal model; rather, they reflect scenarios in which the structured baseline already captures the majority of the relevant pricing signal. In such contexts, the marginal contribution of image and text embeddings is necessarily limited.

A qualitative review of these least-benefit properties reveals a diverse set of asset types and physical characteristics, countering the notion that poor multimodal performance arises from visually “boring” or homogeneous housing stock. Instead, several of the lowest-gain observations correspond to properties whose transactional state inherently suppresses multimodal signal. For example, 233 S Rockingham Ave, among the lowest OOF error-reduction cases, was transacted as vacant land at the time of sale. In this setting, visual and textual descriptors convey little information about architectural quality, interior finishes, or experiential attributes, rendering multimodal embeddings largely uninformative. Under such circumstances, valuation is driven almost entirely by parcel characteristics, zoning potential, and location, all of which are already well-represented in the structured feature set.

This behavior is consistent with a well-calibrated, information-sensitive model rather than an overfit or overly aggressive fusion strategy. Rather than hallucinating signal from sparse or noisy

multimodal inputs, the model effectively “backs off” to structured fundamentals, incurring only a modest loss relative to the structured baseline. From a risk perspective, this bounded downside is a desirable property: it indicates that the multimodal framework enhances valuation where additional information exists, without materially degrading performance where such information is absent.

More broadly, these least-benefit cases underscore an important asymmetry in multimodal valuation. Multimodal features are most valuable in settings where aesthetic, architectural, and experiential factors play a dominant role in price formation which are conditions common in ultra-luxury residential markets but not universal across all transactions. When those factors are muted or nonexistent, structured data alone may suffice, and the model’s ability to recognize this distinction is itself an indicator of robustness.

Taken together, the analysis of least-benefit OOF properties reinforces the interpretation of the multimodal framework as decision-grade rather than purely predictive. The model does not uniformly impose multimodal influence across all observations; instead, it selectively amplifies valuation precision when the underlying data support such inference. This selective behavior aligns with real-world appraisal logic and strengthens confidence in the model’s applicability for high-stakes valuation and risk assessment tasks.

Within-Training-Fold Error Reduction and Regime Saturation

To complement the out-of-fold (OOF) analysis, we examine within-training-fold prediction behavior to assess how model performance changes once the model has been exposed to the underlying feature regimes present in the data. Unlike OOF evaluation, which captures generalization to unseen observations, within-training-fold predictions reflect a setting in which the model has direct access to the data it is evaluated on. As such, this analysis is not intended to measure generalization performance, but rather to contrast model capacity and in-sample representational fidelity.

Model	RMSE (Train)	RMSE (OOF)	MAE (Train)	MAE (OOF)
Ridge	0.360	0.131	0.263	0.095
Full Multimodal (Nonlinear)	0.012	0.019	0.008	0.012

Table 13: Training vs Out-of-Fold Error Comparison (Log Price)

Table 13 compares within-training-fold and OOF error metrics across models. The nonlinear multimodal model achieves near-zero training error (RMSE ≈ 0.012, MAE ≈ 0.008 in log-price space), reflecting its high capacity to closely approximate observed data. In contrast, the structured linear baseline exhibits substantially higher training error (RMSE ≈ 0.36, MAE ≈ 0.26), indicating a limited ability to capture nonlinear relationships even in-sample.

Importantly, while the nonlinear model fits the training data extremely closely, its OOF error remains meaningfully higher, indicating that this expressive capacity does not translate into

trivial memorization. Instead, the model maintains strong generalization performance while leveraging its flexibility to capture complex feature interactions.

Address	Sale Price ($)	Ridge Error (Train)	Nonlinear Error (Train)	Δ Error
1601 San Onofre Dr, Los Angeles	83,000,000	0.0103	0.0402	-0.0299
1155 Angelo Dr, Los Angeles	43,192,890	0.0049	0.0254	-0.0205
1442 Tanager Way, Los Angeles	27,758,104	0.0018	0.0200	-0.0182
1181 N Hillcrest Rd, Beverly Hills	79,277,145	0.0032	0.0188	-0.0156
301 N Carolwood Dr, Los Angeles	109,364,614	0.0081	0.0179	-0.0098
822 Sarbonne Rd, Los Angeles	77,292,540	0.0051	0.0143	-0.0092
1130 Schuyler Rd, Beverly Hills	60,762,469	0.0164	0.0244	-0.0080
33064 Pacific Coast Hwy, Malibu	65,796,938	0.0013	0.0091	-0.0078
457 Bel Air Rd, Los Angeles	63,540,114	0.0008	0.0077	-0.0070
350 N Carolwood Dr, Los Angeles	100,152,312	0.0057	0.0122	-0.0066

Table 14: Properties with the smallest multimodal error reduction within training fold prediction.

Table 14 reports the properties with the smallest multimodal error reduction within training folds. In contrast to the OOF setting, where multimodal features often provide substantial corrective signal, the magnitude of incremental improvement within training is modest relative to the overall fit achieved by the nonlinear model. This does not reflect a convergence between

structured and multimodal performance, but rather the fact that the nonlinear model has already internalized much of the relevant signal during fitting.

Second, properties that exhibited large improvements in the OOF analysis frequently appear among the least-benefit cases within training folds. This pattern is not due to a disappearance of multimodal value, but rather reflects the difference between in-sample fit and out-of-sample correction. In the OOF setting, multimodal features provide additional information when the model must generalize to unseen combinations of architectural, textual, and visual attributes. Within training folds, however, these relationships have already been directly learned, reducing the incremental contribution of additional feature modalities at the margin.

Third, the composition of the least-benefit training-fold properties is heterogeneous, spanning multiple neighborhoods, price tiers, and architectural profiles. This diversity suggests that low marginal multimodal gain within training folds should not be interpreted as an absence of experiential or aesthetic value, but rather as evidence that such value has already been captured within the model's learned representation.

Taken together, the within-training-fold results highlight a key distinction between model capacity and generalization. The nonlinear multimodal model achieves substantially better in-sample fit than the structured baseline, reflecting its ability to capture complex interactions across structural, textual, and visual features. At the same time, the persistence of a gap between training and OOF performance indicates controlled generalization rather than instability.

From a modeling perspective, these findings suggest that multimodality contributes both increased representational flexibility and improved out-of-sample predictive accuracy. Its largest practical value, however, emerges under conditions of uncertainty, where the model must infer value from previously unseen or weakly structured combinations of features.

This distinction is critical for downstream decision-making contexts. In practice, valuations are performed under partial information and distributional shifts, conditions more closely approximated by the OOF setting. The training-fold analysis therefore does not diminish the value of multimodality, but instead clarifies where its benefits are most impactful: when the model must interpret previously unseen or weakly structured market signals.

Toward Decision-Grade Valuation

The findings of this study suggest that valuation in ultra-luxury markets may be more appropriately framed as an uncertainty management problem rather than a point-estimation exercise.

Structured-only models treat residual variance largely as irreducible noise or negotiation friction. However, the multimodal framework materially reduces this residual component, indicating that a portion of apparent “noise” reflects systematically omitted perceptual and experiential information. When textual and visual embeddings are incorporated, prediction error declines sharply relative to structured baselines, and residual dispersion narrows across heterogeneous property types.

This reduction in variance has implications beyond predictive performance metrics. In high-price segments, small percentage deviations translate into substantial absolute capital exposure. A 10% deviation at $20 million corresponds to $2 million in pricing uncertainty; at $100 million, it corresponds to $10 million. Even when percentage errors remain constant, dollar-denominated risk scales nonlinearly with transaction size. Reducing residual variance therefore directly alters the reliability of pricing decisions.

Importantly, “decision-grade” does not imply deterministic precision. It implies calibrated uncertainty. A valuation framework becomes decision-grade when it not only produces point estimates but meaningfully constrains the dispersion around those estimates under out-of-sample evaluation. The cross-validated error reductions observed here, combined with analysis of least-benefit and within-training-fold cases, indicate that multimodal augmentation improves generalization stability rather than merely overfitting to structured data.

The contribution of this study is thus not solely that predictions improve, but that uncertainty becomes more explicitly modeled. In ultra-luxury markets characterized by asset heterogeneity and thin transaction frequency, such calibration may be particularly consequential. As transaction sizes increase and property differentiation intensifies, frameworks that internalize perceptual signal alongside structural attributes may provide more robust risk assessment than structured-only approaches.

Parallels to Other High-Stakes Domains

The structural dynamics observed in ultra-luxury residential valuation resemble transitions seen in other high-stakes domains where informal judgment gradually gave way to formalized risk modeling.

In financial markets, early brokerage practices relied heavily on intuition, narrative framing, and experiential judgment. Quantitative risk frameworks did not displace these elements immediately. They emerged when capital concentration and systemic exposure made unexamined variance untenable. Importantly, adoption was not driven by marginal improvements in average prediction accuracy, but by the need to understand tail behavior and failure modes.

A similar distinction is relevant here. The multimodal framework developed in this study should not be interpreted solely as a more precise pricing engine. Its contribution lies in its capacity to reduce residual variance and expose where structured-only assumptions are incomplete. In doing so, it shifts valuation from a point-estimate exercise toward a risk-aware diagnostic framework.

The analogy is not perfect; ultra-luxury housing lacks the liquidity and systemic interconnectedness of financial markets. Nevertheless, both contexts illustrate a common pattern: as scale and heterogeneity increase, informal expertise becomes insufficient without formal mechanisms for uncertainty calibration. Multimodal modeling represents one such mechanism within residential valuation.

Implications and Future Directions

The framework developed in this study opens several avenues for further investigation, both methodological and structural.

First, the current multimodal implementation operates at the property level. Future work could extend this approach to time-aware modeling, incorporating temporal price dynamics and macroeconomic regimes into the feature space. Ultra-luxury markets are particularly sensitive to liquidity cycles, capital flows, and policy environments. Integrating time-aware cross-validation or rolling-window training may further clarify how multimodal signal strength varies across market conditions.

Second, deeper exploration of uncertainty quantification is warranted. While this study demonstrates substantial reduction in cross-validated residual variance, future research could examine prediction interval calibration, tail-risk behavior, and heteroskedastic error structures. Understanding not only average error reduction but distributional stability under extreme conditions would enhance the risk-aware interpretation of valuation outputs.

Third, interpretability at the embedding level remains an open problem. SHAP analyses reveal the aggregate influence of multimodal signals, but future work could investigate clustering of image and text embeddings to identify latent aesthetic regimes or architectural typologies. Such structure could provide insight into how perceptual segmentation interacts with pricing dynamics in heterogeneous markets.

Beyond these methodological extensions, the broader implication of this study concerns the institutional role of valuation itself. As asset heterogeneity increases and capital allocation decisions grow more concentrated, valuation systems may increasingly be evaluated not solely by point prediction performance but by their ability to surface structural risk and regime sensitivity. A multimodal, diagnostics-oriented framework does not eliminate judgment; it formalizes and disciplines it. Future development may therefore lie not only in refining predictive accuracy, but in embedding such systems within decision workflows that treat valuation as an exercise in structured uncertainty management rather than narrative reinforcement.

In this sense, the present work should be understood as a first step toward a more explicit integration of perceptual data, statistical learning, and risk diagnostics in ultra-luxury residential markets. Whether adopted in practice or extended in academic settings, the central contribution is conceptual as much as technical: valuation can be modeled as a problem of representational completeness and uncertainty compression. The trajectory forward lies in refining, stress-testing, and operationalizing that premise.

References

1. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics.
2. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. NeurIPS.
3. Radford, A., et al. (2021). Learning transferable visual models from natural language supervision. ICML.
4. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. EMNLP.