
HouseTS: A Large-Scale, Multimodal Spatiotemporal U.S. Housing Dataset

Shengkun Wang, Yanshen Sun, Fanglan Chen, Linhan Wang

Naren Ramakrishnan, Chang-Tien Lu, Yinlin Chen

Virginia Tech

Abstract

Accurate house-price forecasting is essential for investors, planners, and researchers. However, reproducible benchmarks with sufficient spatiotemporal depth and contextual richness for long-horizon prediction remain scarce. To address this, we introduce HouseTS—a large-scale, multimodal dataset covering monthly house prices from March 2012 to December 2023 across 6,000 ZIP codes in 30 major U.S. metropolitan areas. The dataset includes over 890K records, enriched with points of Interest (POI), socioeconomic indicators, and detailed real-estate metrics. To establish standardized performance baselines, we evaluate 14 models, spanning classical statistical approaches, deep neural networks (DNNs), and pretrained time-series foundation models. We further demonstrate the value of HouseTS in a multimodal case study, where a vision-language model extracts structured textual descriptions of geographic change from time-stamped satellite imagery. This enables interpretable, grounded insights into urban evolution. HouseTS is hosted on Kaggle, while all preprocessing pipelines, benchmark code, and documentation are openly maintained on GitHub to ensure full reproducibility and easy adoption.

1 Introduction

Accurate house-price prediction is vital for investors, policy makers, and researchers. However, most existing studies rely on narrow data sources, such as past sales or basic demographic counts, and often focus on individual properties without considering broader spatial and temporal patterns [1–6]. Some recent works introduce multimodal inputs like satellite imagery or points of interest, but typically treat them as static features and fail to capture long-term dynamics [7, 8]. While survey papers provide useful overviews of data and methods [9, 10], few offer an open and standardized dataset that reflects the full complexity of housing markets, including both physical and socioeconomic contexts. In addition, many existing time-series datasets in this domain are either too coarse in granularity or overly focused on high-frequency signals, making them unsuitable for long-term, regional housing analysis [11]. This lack of comprehensive, well-aligned, and multimodal resources limits the development, evaluation, and interpretability of forecasting models. A dataset that combines long-term temporal coverage, rich contextual variables, and consistent preprocessing across diverse geographies is therefore urgently needed. Our work directly addresses this gap.

While data limitations pose one major challenge, modeling approaches in house price prediction also face constraints. Many existing studies still rely on statistical techniques or

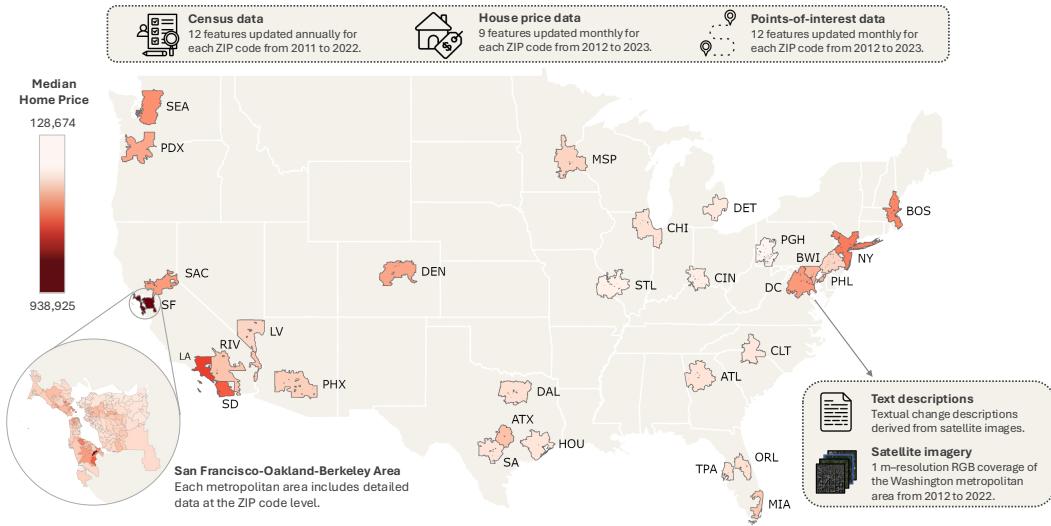


Figure 1: Overview of the HouseTS dataset: spatial coverage across 30 U.S. metropolitan areas and its multimodal data components.

35 conventional machine learning models [12–15]. In parallel, the broader time-series forecasting
 36 community has made significant progress with deep architectures, including Transformer-
 37 based models and large pretrained models designed for sequential data. However, recent
 38 research has raised concerns about their robustness and generalizability. Although these
 39 models perform well on standard benchmarks such as M4 [16], their accuracy often declines
 40 on domain-specific or real-world datasets [17–20]. To better understand their effectiveness in
 41 the housing context, we use our newly curated dataset to rigorously evaluate and compare
 42 deep neural networks and time-series foundation models against traditional baselines, with a
 43 focus on long-term, multivariate house price prediction.

44 To standardize the use of multimodal data and support further research on housing markets
 45 and socioeconomic dynamics, we release a benchmark dataset tailored for long-term urban
 46 house price prediction. The dataset covers more than 6,000 ZIP codes across 30 major
 47 U.S. metropolitan areas and spans the period from 2012 to 2023, as shown in Figure 1. It
 48 integrates detailed monthly and annual house price records, census-based socioeconomic
 49 indicators, neighborhood amenities represented as points of interest, and 1-meter resolution
 50 satellite imagery. We evaluate models under both univariate and multivariate settings, across
 51 multiple input and forecasting horizons. For statistical and machine learning models, we
 52 apply PCA for dimensionality reduction before training. Deep neural networks are tested
 53 with various loss functions and hyperparameter settings, while time-series foundation models
 54 are evaluated in both zero-shot and fine-tuned modes. We also present a multimodal case
 55 study using a vision-language model that generates textual descriptions from satellite image
 56 sequences, allowing us to assess its ability to detect real geographic changes.

57 To address the limitations outlined above, we present HouseTS, a benchmark dataset and
 58 evaluation suite designed for urban house price-related tasks. To the best of our knowledge,
 59 HouseTS offers the widest spatial and temporal coverage, the richest set of modalities,
 60 and the most comprehensive collection of baseline models in this domain (see Table 1).
 61 Specifically, our contributions are threefold: (1) we introduce a large-scale, multimodal
 62 house-price time-series dataset that enables multiple tasks such as forecasting, imputation,
 63 and classification; (2) we establish a benchmark comparing a wide range of models, including
 64 classical models, deep neural networks, and foundation models across multiple horizons;
 65 (3) We conduct in-depth data analysis—including statistical summaries, modality ablation
 66 studies, and a multimodal case study—to illustrate how each modality contributes to modeling
 67 spatiotemporal dynamics.

68 **2 Related Work**

Data source & Research paper	Tabular	Image	Text	Time stamp	Frequency	Horizon	Spatial unit	Observations	Model types
House Sales in King County [1]	✓	X	X	✓	Daily	1 year	Property	21.6K	Stat,ML
Housing Price in Beijing [2]	✓	X	X	✓	Daily	9 years	Property	319K	Stat
Boston Housing Dataset [3]	✓	X	X	-	-	-	Property	0.5K	Stat
Airbnb [21]	✓	X	X	-	-	-	Property	142K	Stat,DNN
OpenStreetMap [8]	✓	✓	✓	-	-	-	Property	470K	ML
Google Map [7]	✓	✓	X	-	-	-	Region	111K	DNN
International House Price Database [22]	✓	✓	X	✓	Quarterly	46 years	Country	4.4K	Stat
FHFA HPI [23]	✓	✓	X	✓	Quarterly	49 years	State	9.3K	Stat
Redfin [8]	✓	X	X	-	-	-	Property	125K	ML
Zillow [4]	✓	X	X	-	-	-	Property	1905K	Stat
HouseTS (Ours)	✓	✓	✓	✓	Daily Monthly	5 years 11 years	Property Region	890K	Stat,ML,DNN/Foundation

Table 1: Comparison of HouseTS with prior house-price datasets and related studies. Left: data source modalities, indicating the types of raw inputs available. Right: how these datasets have been used in past research, including temporal setup, spatial scope, data scale, and model types.

69 House price prediction is a critical task in urban analytics, economics, and real estate planning.
70 Traditional approaches rely on both region-level indicators such as income, unemployment,
71 and housing supply [12, 24], and property-level features including transaction histories,
72 physical attributes, and neighborhood amenities [5, 6]. While these methods offer useful
73 insights, they are often limited by narrow geographic scope and short time spans, making
74 them inadequate for modeling long-term trends and spatial variation. More data types,
75 such as satellite imagery [7, 25], environmental conditions [26], and points of interest [27],
76 offer richer contextual information for prediction. However, these sources differ in spatial
77 resolution, update frequency, and structure, which makes them difficult to integrate into a
78 unified modeling framework.

79 **A variety of open-source datasets** have been proposed for house price research, of-
80 fering either property-level features (e.g., physical attributes, transaction histories, and
81 neighborhood amenities) [28–31] or aggregated indices for broader market trends [32–35].
82 Building upon these resources, house price forecasting has emerged as an active research
83 area, employing a range of techniques at both the individual property level [5, 6] and the
84 regional level [12, 24]. Diverse sources of information have been explored, including real estate
85 transaction records [36], textual descriptions [37], environmental conditions [26], satellite
86 imagery [25], census statistics [38], and points of interest data [27] have been explored to
87 enrich modeling capabilities. However, most existing datasets focus on a single city or cover
88 only short time spans, limiting the potential for long-horizon analyses and cross-region
89 comparisons. To address these limitations, we introduce a new dataset spanning March 2012
90 to December 2023, encompassing 6,000 ZIP codes across 30 major U.S. metropolitan areas.
91 By integrating monthly POI from OpenHistoricalMap[39], socioeconomic metrics from the
92 American Community Survey (ACS)[40], and a broad range of housing market features from
93 Zillow Home Value Index[41] and Redfin statistics[42], our dataset achieves unprecedented
94 granularity in capturing both micro- and macro-level housing trends. Such a multi-faceted
95 resource enables robust cross-city comparisons and advanced tasks like long-horizon forecast-
96 ing with spatiotemporal analyses, filling a critical gap in current benchmarks and providing
97 a versatile platform for real-world time-series research. In addition to the tabular data, we
98 also conduct a case study to evaluate whether existing pretrained multimodal models can
99 extract useful information from satellite images [43] to assist in house price prediction.

100 **Housing price models** rarely adopt recent time-series forecasting methods developed
101 in top-tier machine learning conferences. Although models based on Transformers and
102 pretrained architectures have shown strong results on standard benchmarks, several studies
103 have questioned their robustness and generalization to domain-specific datasets [17–20].
104 To better understand their effectiveness in the housing domain, we evaluate a diverse set
105 of forecasting models on our dataset, including classical statistical baselines, deep neural
106 networks, and state-of-the-art pretrained models. Commonly used datasets as Electricity [44],
107 Traffic [45], and Weather [46]—typically focus on narrower domains and lack the spatial,
108 economic, and temporal complexity found in housing markets. To address this gap and
109 support both the housing and time-series communities, we introduce the HouseTS dataset
110 and benchmark, and systematically evaluate three broad categories of forecasting methods:

- *Statistical approaches and classical machine learning methods:* Statistical models such as ARIMA [47] and VAR [48], as well as traditional machine learning algorithms like Random Forests [49] and XGBoost [50], have been widely used in house price forecasting. These methods are often favored for their interpretability, ease of implementation, and relatively low computational cost. However, their ability to model complex temporal patterns and interactions between heterogeneous data sources remains limited.
- *Deep learning models:* Deep learning methods extend forecasting capabilities by capturing complex temporal dependencies and nonlinear patterns. Recurrent architectures such as RNN [51] and LSTM [52] have been used in various housing-related studies, though they can struggle with long sequences and high-dimensional inputs. More recent models like DLinear [18] and TimeMixer [53] use multi-layer perceptrons for efficient time-series modeling with reduced complexity. Transformer-based architectures, including Informer [54], Autoformer [55], FEDformer[56]and PatchTST [57], introduce innovations across several dimensions. These include point-wise, patch-wise, and variate-wise tokenization schemes, encoder-only or encoder-decoder structures, and alternative attention mechanisms such as ProbSparse attention and Auto-Correlation. However, these models can be sensitive to hyperparameters and may not generalize well without careful adaptation to domain-specific characteristics.
- *Pretrained time-series foundation model:* Recent work has proposed large pretrained models for time-series forecasting, including Chronos [58] and TimesFM [59]. These models are trained on broad collections of time-series data and support zero-shot or few-shot forecasting across different domains. In principle, they offer strong generalization and can incorporate heterogeneous signals such as prices, economic indicators, and even text descriptions. Their modular design enables rapid adaptation without task-specific architectures. However, these models rely on extensive pretraining, are computationally expensive to fine-tune, and may struggle with domain shift when applied to datasets that differ significantly from their original training distribution.

138 3 HouseTS Dataset

139 This section describes the construction of the HouseTS dataset, including data sources,
140 preprocessing procedures, and basic analyses. We provide both raw and cleaned versions of
141 the data, along with reproducible code and visualization notebooks for further exploration.
142 The dataset covers 6,000 ZIP codes across 30 major U.S. metropolitan areas from 2012 to
143 2023 and includes four primary components: house price records, socioeconomic indicators,
144 points of interest, and satellite imagery. Although the benchmark in this paper focuses on
145 house price forecasting, the dataset is also suitable for broader socioeconomic analysis due
146 to its rich coverage of regional amenities and demographic features at the ZIP-code level.
147 The data summary statistics can be found in Table 5.

148 **Points of Interest** capture the availability and density of local amenities within each ZIP
149 code. We collect monthly POI data from March 2012 to December 2023 using the Open
150 Historical Street Network API [60]. Categories include *banks*, *buses*, *hospitals*, *malls*, *parks*,
151 *restaurants*, *schools*, *stations*, and *supermarkets*. For each ZIP code, geographic boundaries
152 were defined using *Geopy*'s Nominatim geocoder, and bounding boxes were used to query
153 POI on a monthly basis. POI data is aggregated as counts per category and timestamp. To
154 handle occasional missing values, we apply a three-stage imputation process. First, we use
155 forward-fill and backward-fill within each ZIP code to fill short gaps. Second, missing values
156 are replaced with the median for that ZIP code across the full time range. Finally, if all
157 values are missing for a ZIP, the overall median across all regions is used. Structural zeros
158 (e.g., truly no hospitals in a ZIP code) are preserved.

159 **Census Data** was collected from the ACS using the U.S. Census Bureau API[40], covering
160 the years 2011 to 2022. This dataset includes annual ZIP-code-level estimates for key
161 demographic and socioeconomic variables such as *Total Population*, *Median Age*, *Per Capita*
162 *Income*, *Total Families Below Poverty*, *Total Housing Units*, *Median Rent*, *Median Home*
163 *Value*, *Total Labor Force*, *Unemployed Population*, *School-Age Population*, *School Enrollment*,
164 and *Median Commute Time*. ZIP codes were mapped to their corresponding state FIPS

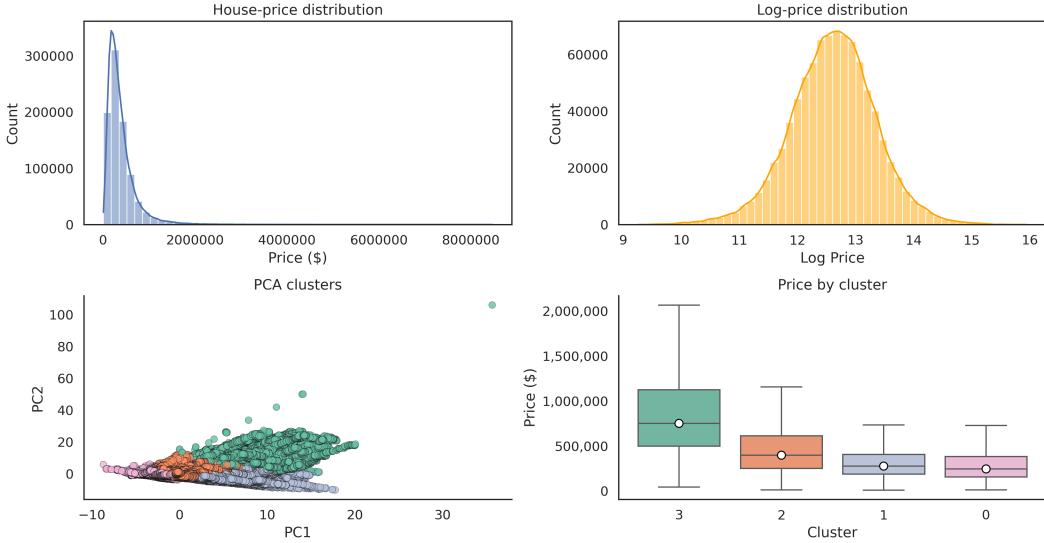


Figure 2: Exploratory analysis of the HouseTS dataset. Top row shows the skewed distribution of raw house prices and the approximately normal distribution after log transformation. Bottom row presents a PCA-based visualization of price clusters and corresponding price distributions per cluster.

165 codes¹ to ensure accurate data extraction. Since ACS updates are annual, we align each
 166 year’s Census data with the following year’s house price records, using 2011–2022 Census
 167 features to predict 2012–2023 prices. This forward-shift ensures that all models operate
 168 under a realistic, temporally valid setup. To address missing values—such as unavailable
 169 estimates for certain ZIPs in some years—we apply the same three-stage imputation strategy
 170 used for POI. Additionally, the ACS API encodes missing values with placeholder negatives,
 171 which we coerce to zero prior to imputation. This process preserves valid zeros and removes
 172 invalid values while ensuring continuity across the dataset.

173 **Historical House Price-Related Features** were collected from two major real estate
 174 platforms: Zillow and Redfin. For both sources, we focus exclusively on the “all residential”
 175 property category to ensure consistency across markets. From Zillow, we obtained monthly
 176 ZIP-code-level home price estimates via the Zillow Home Value Index (ZHVI), which provides
 177 smoothed and seasonally adjusted median price data. Redfin supplies a broader set of monthly
 178 housing market indicators at the ZIP-code level. We include features such as *Median Sale*
 179 *Price*, *Median List Price*, *Median Price per Square Foot*, *Median List Price per Square Foot*,
 180 *Homes Sold*, *Pending Sales*, *New Listings*, *Inventory*, *Median Days on Market*, *Average Sale-to-List Ratio*,
 181 *Share Sold Above List*, and *Share Off Market Within Two Weeks*. Although Redfin also provides derived metrics such as month-over-month and year-over-year growth
 182 rates, we drop these columns to avoid inaccuracies introduced by imputation. To align with
 183 the prediction task, Redfin features are used as leading indicators to inform subsequent Zillow
 184 price predictions. This design reflects realistic market forecasting conditions and avoids label
 185 leakage. The combined dataset spans January 2012 to December 2023 and is aggregated
 186 monthly. Missing values are handled using the same three-stage imputation process applied
 187 to other variables: forward- and backward-fill within each ZIP code, followed by ZIP-level
 188 medians, and fallback to the national median if necessary. Negative placeholders are first
 189 converted to zero and then reprocessed. These housing indicators form the core predictive
 190 target and serve as key inputs in our benchmark evaluation.

192 **Satellite Images** were sourced from the National Agriculture Imagery Program (NAIP)
 193 through Google Earth Engine, with one RGB image per ZIP code per year from 2012 to 2022.
 194 Each image has a spatial resolution of 1 meter and represents a composite of aerial views

¹Federal Information Processing Standard codes are standardized two-digit numerical identifiers used by the U.S. government to uniquely designate each state.

195 within a 200-meter buffer around the ZIP-code boundary. The dataset spans a wide range of
196 geographic contexts, including dense urban cores, suburban developments, and rural areas
197 with sparse infrastructure. Figure 7 illustrates the visual diversity captured in the imagery.
198 Because NAIP collection frequency and coverage vary by state, we focus our multimodal
199 experiments on the Washington metropolitan area, where image availability is consistent
200 across the full time span. Missing tiles for certain years or locations are not interpolated, in
201 order to preserve the integrity of the raw spatial signals. These timestamped images are later
202 used in conjunction with vision-language models to extract structured textual summaries of
203 geographic change, supporting spatiotemporal interpretation of housing price trends. These
204 timestamped images are later processed using a vision–language model to generate textual
205 descriptions of geographic change. This allows structured extraction of spatial information
206 over time and enables interpretable, language-based visualization of urban development. We
207 describe this process in detail in Section 5.

208 To normalize skewed distributions and stabilize learning, we apply a logarithmic transfor-
209 mation to all continuous variables in the POI, Census, and price datasets. As illustrated
210 in Figure 2, raw house prices are heavily right-skewed, while the log-transformed values
211 approximate a Gaussian distribution. Accordingly, we model all targets in the log domain
212 throughout this paper and apply the inverse transformation only when reporting results in
213 original dollar terms.

214 4 Baseline Evaluation

215 We evaluate a broad set of forecasting models on the HouseTS dataset to establish strong,
216 reproducible baselines. In total, we benchmark 14 models across six input–output configu-
217 rations, spanning traditional statistical approaches, classical machine learning algorithms,
218 deep neural networks, and pretrained time series foundation models.

219 4.1 Baseline evaluation methodology

220 Classical statistical models, including ARIMA and VAR, are implemented using the
221 `statsmodels` package [61]. Traditional machine learning models, such as Random Forests
222 and XGBoost, are built with `scikit-learn`[62] and `XGBoost` [50]. Deep learning baselines,
223 including DLinear, TimeMixer, and Informer, are reproduced using the open-source `TSlib`
224 framework [63]. For pretrained time-series foundation models, we evaluate Chronos and
225 TimesFM, both of which are originally designed for univariate forecasting. To support
226 multivariate prediction, we append a lightweight linear projection layer to map outputs to
227 the desired forecasting dimension. For univariate tasks, we use the original author-provided
228 implementations without modification. All models share a consistent preprocessing pipeline.
229 We apply only minimal cleaning, without data augmentation or resampling. Core model
230 architectures are left unchanged, and only lightweight wrappers are added to match our fea-
231 ture layout. As a result, the reported performance is conservative, and further improvements
232 are likely with hyperparameter tuning or domain-specific enhancements.

233 We apply principal component analysis (PCA) to all statistical and machine learning baselines
234 to reduce dimensionality and stabilize training. The original ZIP-level multivariate time
235 series is standardized and projected onto a fixed number of principal components. As shown
236 in Table 6, these components summarize broad patterns such as regional infrastructure,
237 amenity density, and socioeconomic status. Projecting into this reduced space preserves
238 key structural signals while mitigating noise and multicollinearity in the original feature
239 set. Statistical models such as ARIMA and VAR are then trained on the transformed
240 series: ARIMA fits a separate univariate model to each component, while VAR models all
241 components jointly. Tree-based machine learning models, including Random Forest and
242 XGBoost, are trained on lagged PCA features using a direct multi-output strategy, where
243 each model maps a fixed-length input window to the full forecast horizon.

244 Deep learning models are trained by minimizing the mean squared error (MSE) between
245 predicted and observed log-transformed prices. MSE is widely adopted in recent multivariate
246 forecasting work (e.g., Informer, PatchTST, DLinear) due to its simplicity, convexity, and
247 compatibility with gradient-based optimizers such as Adam. It also corresponds to the

248 negative log-likelihood under a Gaussian assumption, which becomes appropriate after log
 249 transformation of the target variable.

250 For pretrained foundation models such as Chronos and TimesFM, we follow the fine-tuning
 251 protocols provided by the original authors. Both models are fully fine-tuned on our dataset.
 252 TimesFM is optimized using the Adam optimizer with a learning rate of 1e-4 over 10 epochs,
 253 employing quantile loss. Chronos is fine-tuned for 2000 steps with a learning rate of 1e-5
 254 using cosine annealing, and uses cross-entropy loss due to its patch-token formulation. To
 255 meet TimesFM's input constraints, all series are zero-padded to a length of 32. For both
 256 models, the checkpoint with the lowest validation loss is chosen for final evaluation.

257 To reduce the impact of outliers and normalize the highly skewed price distribution, we
 258 apply a logarithmic transformation to all target values prior to training and evaluation. This
 259 makes percentage-based deviations comparable across regions and price levels. The primary
 260 evaluation metric is root mean squared error in the log domain (LogRMSE), which penalizes
 261 proportional errors evenly and supports stable optimization. As a secondary diagnostic, we
 262 report mean absolute percentage error (MAPE), which measures relative deviation on the
 263 original scale. We exclude MAE and RMSE, as they disproportionately weight high-end
 264 markets and distort model comparisons.

265 4.2 Baselines evaluation results

Window size →	{6,3}		{6,6}		{6,12}		{12,3}		{12,6}		{12,12}	
	Model ↓	Log-RMSE	MAPE	Log-RMSE								
Repeat	12.993	1.000	12.995	1.000	12.999	1.000	13.017	1.000	13.019	1.000	13.022	1.000
VAR	0.0940	0.0877	0.1073	0.0985	0.1437	0.1256	0.1094	0.1000	0.1363	0.1195	0.4835	0.8442
ARIMA	0.1340	0.1222	0.1560	0.1387	0.1842	0.1601	0.1276	0.1171	0.1504	0.1340	0.1873	0.1615
RandomForest	0.1505	0.1365	0.1703	0.1516	0.2121	0.1840	0.1481	0.1345	0.1668	0.1488	0.2078	0.1804
XGBoost	0.1477	0.1341	0.1730	0.1536	0.2048	0.1777	0.1414	0.1287	0.1660	0.1471	0.2033	0.1762
RNN	0.1254	0.0788	0.1246	0.0810	0.1282	0.0882	0.1219	0.0824	0.1197	0.0807	0.1282	0.0914
LSTM	0.1236	0.0838	0.1258	0.0844	0.1327	0.0919	0.1252	0.0846	0.1263	0.0879	0.1325	0.0964
DLinear	5.4489	0.9942	6.3387	0.9976	6.8727	0.9985	6.7197	0.9984	6.9611	0.9987	7.4813	0.9992
Autoformer	1.1986	0.7657	1.2640	0.6059	1.2366	1.3895	1.4602	0.5376	1.1937	0.7609	1.2306	0.7009
PatchTST	7.9983	1.0859	8.4042	1.0839	7.6521	1.0797	7.7163	1.1790	7.8870	1.1537	7.2245	1.2020
FEDformer	2.0208	0.7052	1.4250	0.5464	1.8235	0.6443	3.3039	1.2273	1.6808	0.6047	1.7453	0.6452
Informer	0.1568	0.1073	0.1729	0.1261	0.1736	0.1303	0.1652	0.1197	0.1879	0.1560	0.1622	0.1129
TimeMixer	1.0085	0.6244	0.8827	0.5727	0.9585	0.5998	0.9114	0.6375	0.7405	0.5243	0.7238	0.5154
TimesFM _{zero}	0.2881	0.0780	0.3223	0.0793	0.2976	0.0876	0.0434	0.0303	0.0547	0.0394	0.0734	0.0553
TimesFM	0.0562	0.0214	0.0381	0.0263	0.0684	0.0518	0.0327	0.0178	0.0717	0.0422	0.0693	0.0423
Chronos _{zero}	0.0946	0.0597	0.1238	0.0846	0.1642	0.1216	0.0522	0.0372	0.0679	0.0499	0.0935	0.0719
Chronos	0.0352	0.0220	0.0425	0.0303	0.0709	0.0489	0.0381	0.0259	0.0483	0.0344	0.0772	0.0530

Table 2: Performance comparison of models on multivariate house-price forecasting. The lowest (best) value in each metric column is highlighted.

Window size →	{6,3}		{6,6}		{6,12}		{12,3}		{12,6}		{12,12}	
	Model ↓	Log-RMSE	MAPE	Log-RMSE	MAPE	Log-RMSE	MAPE	Log-RMSE	MAPE	Log-RMSE	MAPE	Log-RMSE
Repeat	12.993	1.000	12.995	1.000	12.999	1.000	13.017	1.000	13.019	1.000	13.022	1.000
AR	0.7742	0.8489	0.5664	0.4971	0.7415	0.8301	0.7547	0.8485	0.5501	0.4885	0.7371	0.8381
ARIMA	0.7612	0.8613	0.5444	0.4719	0.7403	0.8721	0.7868	0.9518	0.5631	0.5247	0.7477	0.9145
RandomForest	0.5762	0.6041	0.8892	1.1975	0.9209	1.3546	0.3743	0.2355	0.9080	1.2620	1.0204	1.4280
XGBoost	0.5543	0.5787	0.8759	1.1226	0.9082	1.3062	0.4076	0.2541	0.8898	1.1886	1.0150	1.4297
RNN	0.1243	0.0783	0.1232	0.0805	0.1267	0.0877	0.1205	0.0818	0.1185	0.0802	0.1268	0.0908
LSTM	0.1236	0.0838	0.1258	0.0844	0.1327	0.0919	0.1252	0.0846	0.1263	0.0879	0.1325	0.0964
DLinear	5.4942	0.9935	6.4043	0.9971	6.9515	0.9983	6.7902	1.0006	7.0400	1.0003	7.5204	1.0009
Autoformer	1.2950	0.9976	1.1163	0.8974	1.4483	0.6683	1.7669	0.6364	1.2632	0.6418	1.1785	0.7056
PatchTST	6.8826	1.3486	7.2218	1.3643	7.6403	1.3830	7.0286	1.4062	7.0904	1.4170	7.0905	1.4223
FEDformer	1.1912	0.7514	1.6941	0.7419	1.4899	0.6581	2.4633	2.5296	1.8173	0.7152	1.8815	0.7564
Informer	0.1033	0.07614	0.0776	0.0663	0.08329	0.0635	0.0744	0.0626	0.0732	0.0599	0.0833	0.0673
TimeMixer	0.9450	0.5961	1.2895	0.7053	1.1748	0.6787	1.0164	0.6184	0.9723	0.6069	1.0396	0.6277
TimesFM _{zero}	0.7864	0.1675	0.8113	0.1721	0.8433	0.1980	0.0931	0.0634	0.1016	0.0710	0.1245	0.0882
TimesFM	0.0132	0.0089	0.0313	0.0222	0.0610	0.0422	0.0174	0.0123	0.0327	0.0222	0.0594	0.0400
Chronos _{zero}	0.0821	0.0436	0.1042	0.0649	0.1376	0.0948	0.0157	0.0264	0.0436	0.0280	0.0661	0.0451
Chronos	0.0335	0.0163	0.0356	0.0163	0.0671	0.0435	0.0211	0.0115	0.0336	0.0216	0.0673	0.0426

Table 3: Univariate forecasting performance using only house-price data. The lowest (best) value in each metric column is highlighted.

266 Table 2 reports the multivariate results for seventeen candidate models, spanning classical
 267 statistics, tree-based learners, a range of specialized neural architectures, and two fine-tuned
 268 foundation models (Chronos and TimesFM). The companion univariate scores, obtained when
 269 only the log-transformed price series are available, are presented in Table 3. Across every

metric column in both tables, the minimum error is achieved by one of the two foundation models, underscoring the benefit of large-scale pre-training followed by light task-specific fine-tuning. TimesFM attains the lowest Log-RMSE on all horizons and delivers the best MAPE for the {6, 6}, {12, 3}, and {12, 12} configurations, whereas Chronos secures the top MAPE at {6, 12}; the pair therefore establishes a clear upper bound for long-range accuracy. Careful examination of the publicly released pre-training corpora confirms that neither model was exposed to house-price or real-estate valuation data, indicating that their gains arise from generic temporal reasoning rather than latent domain leakage.

Traditional statistical baselines (VAR, ARIMA) remain serviceable on the shortest window {6, 3}, yet their errors grow monotonically as the forecast horizon lengthens. Tree-based ensembles (Random Forest, XGBoost) follow a similar trajectory, outperforming the statistical methods in several mid-range settings but ultimately lagging behind the neural approaches. Among bespoke deep-learning architectures, the recurrent families (RNN and LSTM) provide consistently solid—though not leading—performance, while Informer stands out as the most stable Transformer variant: it maintains low error on every horizon and is the only non-foundation model that approaches the foundation benchmarks. By contrast, Autoformer and FEDformer deteriorate sharply on longer windows, and scale-sensitive designs such as DLinear and PatchTST exhibit pronounced instability, with spuriously large Log-RMSE values that betray a poor fit to the multivariate price dynamics. In terms of computational cost, classical statistical methods and tree-based learners terminate quickest, recurrent networks and lightweight linear baselines occupy a middle tier, Transformer architectures demand substantially more computation, and the fine-tuned foundation models—Chronos and TimesFM—incur the longest training times.

5 A Multimodal Case Study for the Washington Metropolitan Area



Figure 3: Illustrative example from our multimodal case study: (top) a time-ordered sequence of satellite tiles for ZIP code 22305; (bottom) the geo-textual description produced by our multimodal large model; (right) the multimodal prediction pipeline.

To demonstrate the multimodal potential of the HouseTS dataset, we conduct a case study focused on the Washington D.C.–Maryland–Virginia (DMV) metropolitan area, where yearly high-resolution satellite imagery is consistently available. This experiment showcases how visual data, when combined with house price records, can be used to extract structured geographic insights and support interpretable spatiotemporal analysis.

300 For each of the 308 ZIP codes in the DMV area, we align a 10-year sequence of satellite
 301 images with annual house price trends. We then reserve the final year as a prediction
 302 target. Using a vision–language model, we convert each ZIP’s image sequence into a textual
 303 summary of observable changes—such as development density, land use transformation, or
 304 infrastructure expansion. Each annual satellite image is preprocessed into a standardized 512
 305 \times 512 RGB tile centered on the ZIP code boundary. To ensure consistency across inputs, we
 306 fix the spatial scale and cropping strategy for all ZIP codes. Detailed prompting strategies for
 307 both text generation and multimodal forecasting can be found in Figure 6. These generated
 308 descriptions serve as an intermediate modality, capturing long-term urban evolution in a
 309 format that can complement numerical features. An illustration of this pipeline is shown in
 310 Figure 3 (right).
 311 To obtain the text modality data, we apply GPT-o3 to each ZIP code’s image sequence,
 312 prompting it to generate a summary of observed changes and local characteristics. These
 313 descriptions capture macro trends (e.g., urban expansion), micro-level developments (e.g., new
 314 buildings or roadways), and static features (e.g., green space density). The text information
 315 shown at Figure 3 (bottom) reveal that the advanced multimodal model captures geo-spatial
 316 change most faithfully. While its ability to predict house prices from images alone remains
 317 limited, GPT-o3 still extracts usable geographic cues from multi-year satellite sequences,
 318 confirming that visual information could potentially enrich the price-only baseline. Minor
 319 hallucinations do occur, but they do not materially affect the overall trend detection or the
 320 geographic stratification insights observed.
 321 To demonstrate the multimodal utility of HouseTS, we conduct three evaluations within
 322 the DMV subset. First, an image-only test examines whether satellite imagery alone can
 323 reveal geographic cues relevant to housing price trends. Second, a temporal forecasting test
 324 evaluates whether historical imagery sequences contain signals predictive of future prices.
 325 Third, a multimodal comparison assesses the impact of augmenting price data with either
 326 raw satellite images or image-derived textual descriptions.
 327 All forecasting experiments, including mul-
 328 timodal ones, are performed using GPT-
 329 4o, which takes as input either the struc-
 330 tured tabular data alone or in combina-
 331 tion with image or text features. Results in Ta-
 332 ble 4 show that while the price-only base-
 333 line remains strongest, incorporating image-
 334 derived text improves performance over raw
 335 imagery in terms of MAPE. This suggests
 336 that translating visual content into struc-
 337 tured descriptions has the potential to enhance
 338 model interpretability and support more
 robust downstream prediction.

<i>Price</i>	✓	✓	✓	✓
<i>Image</i>	✗	✓	✗	✓
<i>Text</i>	✗	✗	✓	✓
Log-RMSE	0.1840	2.2903	3.7376	2.5526
MAPE	0.1568	0.3205	0.3878	0.2520

Table 4: Ablation study results on Price, Image, and Text of GPT-4o.

339 6 Conclusion

340 We introduce HouseTS, a comprehensive multimodal dataset for long-term house price predic-
 341 tion, covering over 6,000 ZIP codes across 30 major U.S. cities over a 10-year span. Compared
 342 to existing datasets, HouseTS provides broader temporal coverage, wider geographic scope,
 343 and richer data modalities—including high-resolution satellite imagery, socioeconomic in-
 344 dicators, neighborhood amenities, and detailed housing price series. We establish a strong
 345 benchmark with 14 baseline models, spanning statistical, machine learning, deep learning,
 346 and foundation models, evaluated under both zero-shot and fine-tuned settings. We further
 347 demonstrate the potential of multimodal large models to capture spatiotemporal patterns
 348 through structured image-to-text pipelines. All preprocessing code, benchmark implemen-
 349 tations, and model outputs are publicly available to ensure reproducibility and facilitate fair
 350 comparisons. Beyond forecasting, HouseTS supports related tasks such as imputation, urban
 351 clustering, and socioeconomic trend analysis, making it a versatile resource for advancing
 352 both methodological and applied research in housing markets and urban analytics.

353 **References**

- 354 [1] CH. Raga Madhuri, G. Anuradha, and M. Vani Pujitha. House price prediction using
355 regression techniques: A comparative study. In *2019 International Conference on Smart
356 Structures and Systems (ICSSS)*, pages 1–5, 2019. doi: 10.1109/ICSSS.2019.8882834.
- 357 [2] Quang Truong, Minh Nguyen, Hy Dang, and Bo Mei. Housing price prediction via
358 improved machine learning techniques. *Procedia Computer Science*, 174:433–442, 2020.
- 359 [3] Abigail Bola Adetunji, Oluwatobi Noah Akande, Funmilola Alaba Ajala, Ololade Oyewo,
360 Yetunde Faith Akande, and Gbenle Oluwadara. House price prediction using random
361 forest machine learning technique. *Procedia Computer Science*, 199:806–813, 2022.
- 362 [4] Qinan Lu, Nienan Cheng, Wendong Zhang, and Pengfei Liu. Disamenity or premium:
363 Do electricity transmission lines affect farmland values and housing prices differently?
364 *Journal of Housing Economics*, 62:101968, 2023.
- 365 [5] Byeonghwa Park and Jae Kwon Bae. Using machine learning algorithms for housing
366 price prediction: The case of fairfax county, virginia housing data. *Expert systems with
367 applications*, 42(6):2928–2934, 2015.
- 368 [6] Winky KO Ho, Bo-Sin Tang, and Siu Wai Wong. Predicting property prices with
369 machine learning algorithms. *Journal of Property Research*, 38(1):48–70, 2021.
- 370 [7] Stephen Law, Brooks Paige, and Chris Russell. Take a look around: using street view
371 and satellite images to estimate house prices. *ACM Transactions on Intelligent Systems
372 and Technology (TIST)*, 10(5):1–19, 2019.
- 373 [8] Yuhao Kang, Fan Zhang, Wenzhe Peng, Song Gao, Jimmeng Rao, Fabio Duarte, and
374 Carlo Ratti. Understanding house price appreciation using multi-source big geo-data
375 and machine learning. *Land use policy*, 111:104919, 2021.
- 376 [9] Margot Geerts, Seppe Vanden Broucke, and Jochen De Weerdt. A survey of methods
377 and input data types for house price prediction. *ISPRS International Journal of
378 Geo-Information*, 12(5):200, 2023.
- 379 [10] Nor Hamizah Zulkifley, Shuzlina Abdul Rahman, Nor Hasbiah Ubaidullah, and Ismail
380 Ibrahim. House price prediction using a machine learning model: a survey of literature.
381 *International Journal of Modern Education and Computer Science*, 12(6):46–54, 2020.
- 382 [11] Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming
383 Jin. Time-moe: Billion-scale time series foundation models with mixture of experts,
384 2024. URL <https://arxiv.org/abs/2409.16040>.
- 385 [12] Shahrzad Ghouchian and Hakan Yilmazkuday. Housing price dynamics within the us:
386 Evidence from zip codes with different demographics. Available at SSRN 3575021, 2024.
- 387 [13] Ali Soltani, Mohammad Heydari, Fatemeh Aghaei, and Christopher James Pettit. Housing
388 price prediction incorporating spatio-temporal dependency into machine learning
389 algorithms. *Cities*, 131:103941, 2022.
- 390 [14] Juan Ramón Rico-Juan and Paloma Taltavull de La Paz. Machine learning with
391 explainability or spatial hedonics tools? an analysis of the asking prices in the housing
392 market in alicante, spain. *Expert Systems with Applications*, 171:114590, 2021.
- 393 [15] Yu Zhang, Dachuan Zhang, and Eric J Miller. Spatial autoregressive analysis and mod-
394eling of housing prices in city of toronto. *Journal of Urban Planning and Development*,
395 147(1):05021003, 2021.
- 396 [16] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m4 com-
397 petition: 100,000 time series and 61 forecasting methods. *International Journal of
398 Forecasting*, 36(1):54–74, 2020.

- 399 [17] Mingtian Tan, Mike Merrill, Vinayak Gupta, Tim Althoff, and Tom Hartvigsen. Are
 400 language models actually useful for time series forecasting? *Advances in Neural*
 401 *Information Processing Systems*, 37:60162–60191, 2024.
- 402 [18] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time
 403 series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*,
 404 volume 37, pages 11121–11128, 2023.
- 405 [19] Christoph Bergmeir. Llms and foundational models: Not (yet) as good as hoped.
 406 *Foresight: The International Journal of Applied Forecasting*, 73, 2024.
- 407 [20] Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I Webb, Rob J Hyndman, and
 408 Pablo Montero-Manso. Monash time series forecasting archive. *arXiv preprint*
 409 *arXiv:2105.06643*, 2021.
- 410 [21] Nicola Camatti, Giacomo di Tollo, Gianni Filograsso, and Sara Ghilardi. Predicting
 411 airbnb pricing: a comparative analysis of artificial intelligence and traditional approaches.
 412 *Computational Management Science*, 21(1):30, 2024.
- 413 [22] Xichen Wang and Qingya Liu. Can the global financial cycle explain the episodes of
 414 exuberance in international housing markets? *Finance Research Letters*, 52:103366,
 415 2023.
- 416 [23] Anastasios G Malliaris, Mary Malliaris, and Mark S Rzepczynski. One man’s bubble is
 417 another man’s rational behavior: comparing alternative macroeconomic hypotheses for
 418 the us housing market. *Journal of Risk and Financial Management*, 17(8):349, 2024.
- 419 [24] Atif Mian and Amir Sufi. House prices, home equity-based borrowing, and the us
 420 household leverage crisis. *American Economic Review*, 101(5):2132–2156, 2011.
- 421 [25] Pei-Ying Wang, Chiao-Ting Chen, Jain-Wun Su, Ting-Yun Wang, and Szu-Hao Huang.
 422 Deep learning model for house price prediction using heterogeneous data analysis along
 423 with joint self-attention mechanism. *IEEE access*, 9:55244–55259, 2021.
- 424 [26] Yixiong Xiao, Xiang Chen, Qiang Li, Xi Yu, Jin Chen, and Jing Guo. Exploring
 425 determinants of housing prices in beijing: An enhanced hedonic regression with open
 426 access poi data. *ISPRS International Journal of Geo-Information*, 6(11):358, 2017.
- 427 [27] Linchuan Yang, Bo Wang, Jiangping Zhou, and Xu Wang. Walking accessibility and
 428 property prices. *Transportation Research Part D: Transport and Environment*, 62:
 429 551–562, 2018.
- 430 [28] Shree. House price prediction dataset, 2018. URL <https://www.kaggle.com/datasets/shree1992/housedata/data>.
- 432 [29] Dan Becker. Melbourne housing snapshot, 2018. URL <https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot>.
- 434 [30] Cam Nugent. California housing prices, 2017. URL <https://www.kaggle.com/datasets/camnugent/california-housing-prices>.
- 436 [31] Anthony Pino. Melbourne housing market, 2017. URL <https://www.kaggle.com/datasets/anthonymino/melbourne-housing-market>.
- 438 [32] Federal Housing Finance Agency. Fhfa house price index (hpi), 2025. URL <https://www.fhfa.gov/data/hpi>. Accessed: 2025-04-08.
- 440 [33] Zillow. Zillow home value index (zhvi), 2025. URL <https://www.zillow.com/research/data/>. Accessed: 2025-04-08.
- 442 [34] Redfin. Redfin housing market data, 2025. URL <https://www.redfin.com/news/data-center/>. Accessed: 2025-04-08.
- 444 [35] Realtor.com. Realtor.com real estate data and market trends, 2025. URL <https://www.realtor.com/research/data/>. Accessed: 2025-04-08.

- 446 [36] Dean De Cock. Ames, iowa: Alternative to the boston housing data as an end of
 447 semester regression project. *Journal of Statistics Education*, 19(3), 2011.
- 448 [37] Lulin Xu and Zhongwu Li. A new appraisal model of second-hand housing prices in
 449 china's first-tier cities based on machine learning algorithms. *Computational Economics*,
 450 57(2):617–637, 2021.
- 451 [38] José-María Montero, Román Mínguez, and Gema Fernández-Avilés. Housing price
 452 prediction: parametric versus semi-parametric spatial hedonic models. *Journal of*
 453 *Geographical Systems*, 20:27–55, 2018.
- 454 [39] OpenHistoricalMap contributors. OpenHistoricalMap. [https://www.](https://www.openhistoricalmap.org/)
 455 openhistoricalmap.org/, 2025. [Accessed: 2025-04-09].
- 456 [40] U.S. Census Bureau. American Community Survey. [https://www.census.gov/](https://www.census.gov/programs-surveys/acs)
 457 programs-surveys/acs, 2025. [Accessed: 2025-04-09].
- 458 [41] Zillow Research. Zillow Housing Data. <https://www.zillow.com/research/data/>,
 459 2025. [Accessed: 2025-04-09].
- 460 [42] Redfin Corporation. Redfin Data Center. [https://www.redfin.com/news/](https://www.redfin.com/news/data-center/)
 461 data-center/, 2025. [Accessed: 2025-04-09].
- 462 [43] U.S. Department of Agriculture. National Agriculture Imagery Program (NAIP).
 463 <https://naip-usdaonline.hub.arcgis.com/>, 2025. Accessed: 2025-04-14.
- 464 [44] UCI Machine Learning Repository. Electricity load diagrams 2011-2014 data set. <https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014>, 2014.
 465 Accessed 2023-XX-XX.
- 466 [45] California Department of Transportation. Caltrans performance measurement system
 467 (pems). <https://pems.dot.ca.gov/>, 2017. Accessed 2023-XX-XX.
- 468 [46] NOAA. National centers for environmental information. <https://www.ncei.noaa.gov/>,
 469 2017. Data often referenced in Lai et al., SIGIR 2018, and other time-series papers.
- 470 [47] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time*
 471 *series analysis: forecasting and control*. John Wiley & Sons, 2015.
- 472 [48] Christopher A Sims. Macroeconomics and reality. *Econometrica: journal of the*
 473 *Econometric Society*, pages 1–48, 1980.
- 474 [49] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.
- 475 [50] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In
 476 *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery*
 477 *and data mining*, pages 785–794, 2016.
- 478 [51] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.
- 479 [52] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*,
 480 9(8):1735–1780, 1997.
- 481 [53] Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y
 482 Zhang, and JUN ZHOU. Timemixer: Decomposable multiscale mixing for time series
 483 forecasting. In *International Conference on Learning Representations (ICLR)*, 2024.
- 484 [54] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and
 485 Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series
 486 forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35,
 487 pages 11106–11115, 2021.
- 488 [55] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition
 489 transformers with auto-correlation for long-term series forecasting. *Advances in neural*
 490 *information processing systems*, 34:22419–22430, 2021.

- 492 [56] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fed-
493 former: Frequency enhanced decomposed transformer for long-term series forecasting.
494 In *International conference on machine learning*, pages 27268–27286. PMLR, 2022.
- 495 [57] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time
496 series is worth 64 words: Long-term forecasting with transformers. In *International*
497 *Conference on Learning Representations*, 2023.
- 498 [58] Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado,
499 Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango,
500 Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint*
501 *arXiv:2403.07815*, 2024.
- 502 [59] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation
503 model for time-series forecasting. In *Forty-first International Conference on Machine*
504 *Learning*, 2024.
- 505 [60] Martin Raifer, Rafael Troilo, Fabian Kowatsch, Michael Auer, Lukas Loos, Sabrina
506 Marx, Katharina Przybill, Sascha Fendrich, Franz-Benjamin Mocnik, and Alexander
507 Zipf. Oshdb: a framework for spatio-temporal analysis of openstreetmap history data.
508 *Open Geospatial Data, Software and Standards*, 4(1):1–12, 2019.
- 509 [61] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling
510 with python. In *9th Python in Science Conference*, 2010.
- 511 [62] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel,
512 P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau,
513 M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python.
514 *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- 515 [63] Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Mingsheng Long, and Jianmin
516 Wang. Deep time series models: A comprehensive survey and benchmark. 2024.

517 **Appendix**

518 **A Feature Analysis**

Feature	Mean	Std	Min	25%	50%	75%	Max
Median Age	36.734	12.437	0.000	34.100	38.500	43.100	91.200
Median Commute Time	9,687.920	8,841.743	0.000	1,945.000	7,830.500	15,013.000	60,956.000
Median Home Value	314,792.244	267,219.362	0.000	143,600.000	250,100.000	412,500.000	2,000,001.000
Median Rent	1,146.686	547.237	0.000	852.000	1,114.000	1,446.000	3,501.000
Per Capita Income	35,253.501	21,555.152	0.000	23,210.000	32,025.000	44,198.000	465,868.000
Total Families Below Poverty	21,457.525	19,554.170	0.000	4,404.000	17,489.000	32,991.000	130,605.000
Total Housing Units	8,714.481	7,588.635	0.000	1,930.000	7,426.000	13,564.000	48,734.000
Total Labor Force	11,455.780	10,429.516	0.000	2,320.000	9,299.000	17,702.000	68,735.000
Total Population	21,802.546	19,794.374	0.000	4,512.000	17,848.000	33,538.000	130,920.000
Total School Age Population	20,998.338	19,008.391	0.000	4,369.000	17,247.000	32,290.000	126,948.000
Total School Enrollment	20,998.338	19,008.391	0.000	4,369.000	17,247.000	32,290.000	126,948.000
Unemployed Population	829.769	954.754	0.000	127.000	538.000	1,192.000	9,735.000
avg_sale_to_list	0.978	0.064	0.000	0.965	0.982	0.998	1.906
bank	13.384	31.045	0.000	0.000	4.000	15.000	447.000
bus	0.670	1.610	0.000	0.000	0.000	1.000	26.000
homes_sold	76.723	76.698	0.000	19.000	55.000	111.000	955.000
hospital	3.506	7.368	0.000	0.000	1.000	4.000	96.000
inventory	77.301	89.042	0.000	20.000	50.000	103.000	1,941.000
mall	1.292	2.752	0.000	0.000	0.000	1.000	45.000
median_dom	61.290	82.220	0.000	26.000	45.000	74.000	7,777.000
median_list_ppsf	231.170	290.120	0.000	116.818	173.143	270.181	143,015.399
median_list_price	422,984.881	1,899,201.111	0.000	199,000.000	320,000.000	499,900.000	999,999,999.000
median_ppsf	223.068	696.724	0.000	110.640	166.094	260.626	366,700.000
median_sale_price	394,102.626	381,548.138	0.000	185,000.000	302,500.000	480,000.000	20,500,000.000
new_listings	92.910	92.696	0.000	24.000	67.000	133.000	1,112.000
off_market_in_two_weeks	0.306	0.239	0.000	0.083	0.295	0.476	1.000
park	48.989	75.719	0.000	5.000	24.000	63.000	926.000
pending_sales	81.471	85.328	0.000	17.000	57.000	119.000	1,374.000
price	391,328.910	344,538.332	10,464.318	189,706.296	305,018.960	479,711.108	8,463,115.592
restaurant	64.993	199.437	0.000	2.000	13.000	50.000	3,409.000
school	48.667	62.302	0.000	7.000	27.000	66.000	560.000
sold_above_list	0.264	0.202	0.000	0.120	0.224	0.375	1.000
station	5.703	16.774	0.000	0.000	0.000	4.000	192.000
supermarket	9.718	19.202	0.000	1.000	4.000	12.000	303.000

Table 5: Descriptive statistics for features.

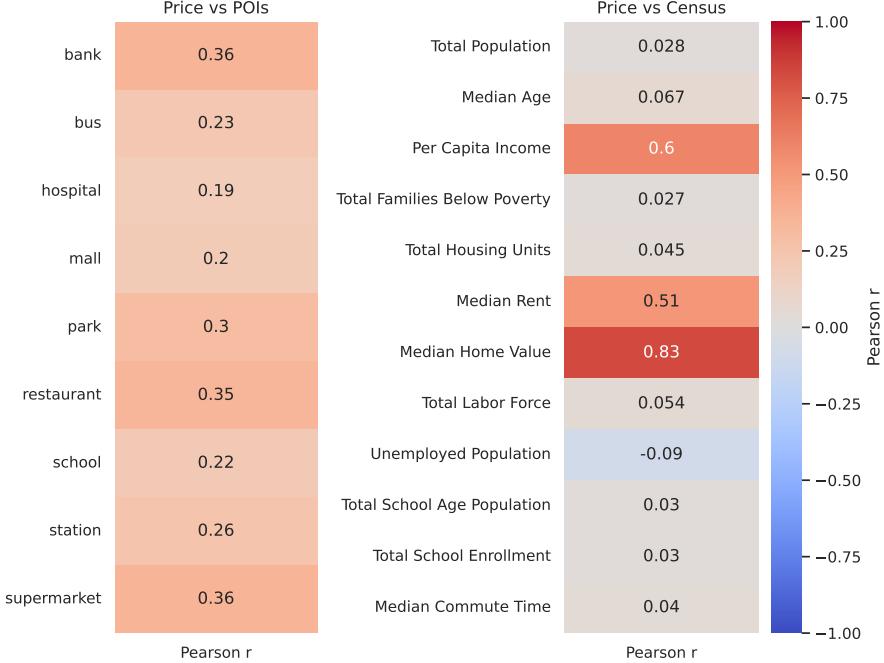


Figure 4: Pearson r correlations between median house price and (left) POI densities and (right) census variables.

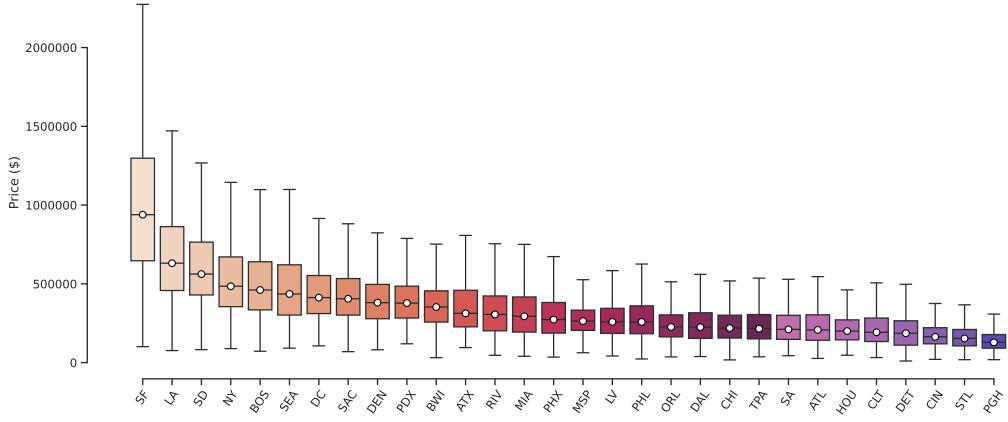


Figure 5: House price distribution across regions covered in the HouseTS dataset. Boxes show interquartile range; whiskers indicate data spread; medians are marked with white dots.

PC	Feature	Loading
PC ₁	Total Housing Units	0.286
	Total Labor Force	0.285
	Median Commute Time	0.282
	Total School Enrollment	0.281
	Total School Age Population	0.281
PC ₂	Restaurant	0.302
	Bank	0.302
	Supermarket	0.287
	Station	0.280
	Park	0.267
PC ₃	Median Rent	0.365
	Per Capita Income	0.353
	Median Home Value	0.336
	Median Age	0.297
	Off-market in two weeks	0.273

Table 6: Absolute top-5 loadings of the first three principal components.

⁵¹⁹ B Prompts for Multimodal Case Study

(a) Generated textual geo-information prompt

```
You are an urban remote-sensing analyst.

### Input
• You will receive  $N$  satellite images (.png) whose filenames follow <YYYY>.png (year order is not guaranteed).

### Tasks
1. Analyse land-use evolution across all images.
2. Decide the overall density (sparse | medium | dense).
3. Decide the overall setting (urban | suburban | rural).
4. Summarise trends  $\leq 60$  words.
5. Provide exactly 5 keywords.
6. List 3-6 notable changes.

### Output (strict JSON)
{ "trend_summary": "...", "keywords": [...], "notable_changes": [...] }
```

(b) Price-only prediction prompt

```
System prompt
You are an experienced real-estate market analyst. I will give you historical year-end median home prices for one U.S. ZIP code. Return ONLY the predicted median home price (number, no $ or commas) for the next year.

User message
Here are the year-end prices:
<YYYY : price lines>

What is your prediction for <next_year>?
```

(c) Image + Price prediction prompt

```
System prompt
You are a real-estate analyst combining satellite imagery and historical price data to forecast median home prices. Given several images (chronological order) and the year-end prices for those same years, return ONLY the predicted price (number, no $ sign or commas) for the next year.

User message
Images: <image-URL list>
Here are the year-end prices:
<YYYY : price lines>

What is your prediction for <next_year>?
```

(d) Image + Text + Price prediction prompt

```
System prompt
You are a real-estate analyst combining satellite images, their semantic description, and historical prices. Predict NEXT year's median home price. Return ONLY a JSON object:
{"price": <number>} (no $ or commas).

User message
Images: <base64-encoded PNG list>
Satellite semantics:
- Summary: <trend_summary>
- Keywords: <keyword_list>
- Notable changes: <notable_changes>

Year-end prices:
<YYYY : price lines>

What is your prediction for <next_year>?
```

Figure 6: Prompt templates for textual geo-information generation and various data-modality forecasting tasks.

520 **C Satellite Image Example**

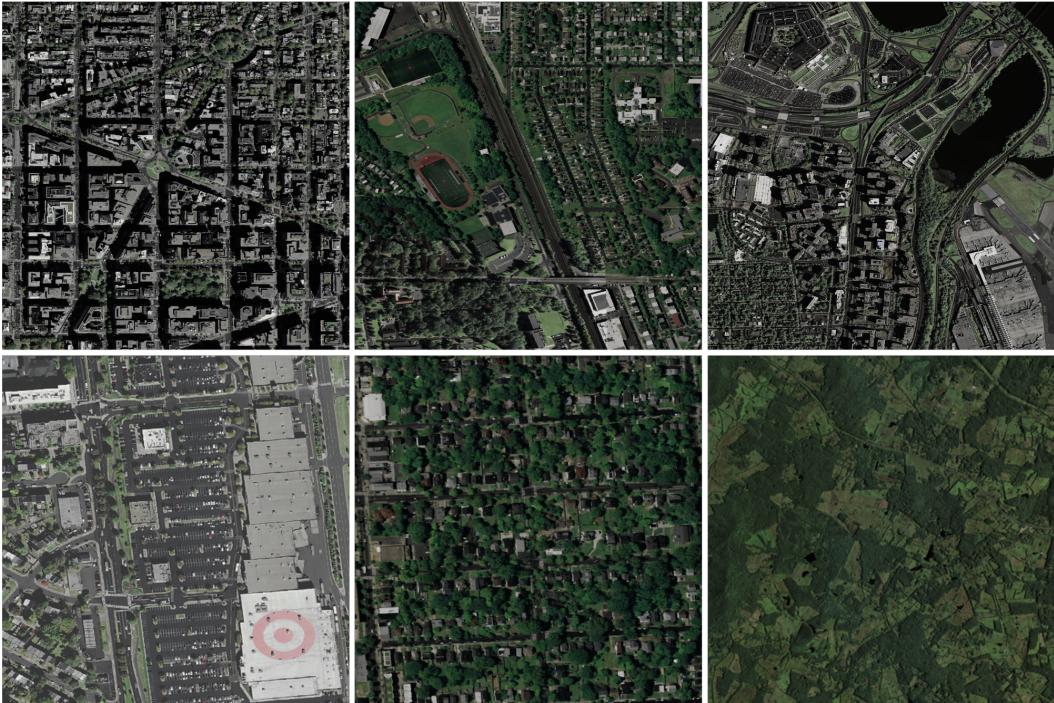


Figure 7: Sample Satellite image, illustrating the dataset’s geographic breadth, from dense downtown blocks and transit-oriented suburbs to big-box commercial strips, leafy single-family grids, and open rural landscapes.

521 **D Limitation and Negative Impact**

522 While HouseTS provides a solid foundation for multimodal housing research, sev-
523 eral limitations remain. Satellite imagery is currently limited to the Washington
524 D.C.–Maryland–Virginia area, and NAIP’s rolling acquisition cycle results in uneven annual
525 coverage. Some missing data may stem from source gaps or collection issues. A potential
526 negative impact is that predictive outputs could be misapplied in policy or financial contexts
527 if used without proper consideration of model uncertainty and data limitations.

528 **NeurIPS Paper Checklist**

529 **1. Claims**

530 Question: Do the main claims made in the abstract and introduction accurately
531 reflect the paper's contributions and scope?

532 Answer: [Yes]

533 Justification: We clearly state our contributions and scope in the abstract.

534 Guidelines:

- 535 • The answer NA means that the abstract and introduction do not include the
536 claims made in the paper.
- 537 • The abstract and/or introduction should clearly state the claims made, including
538 the contributions made in the paper and important assumptions and limitations.
539 A No or NA answer to this question will not be perceived well by the reviewers.
- 540 • The claims made should match theoretical and experimental results, and reflect
541 how much the results can be expected to generalize to other settings.
- 542 • It is fine to include aspirational goals as motivation as long as it is clear that
543 these goals are not attained by the paper.

544 **2. Limitations**

545 Question: Does the paper discuss the limitations of the work performed by the
546 authors?

547 Answer: [Yes]

548 Justification: We address the limitations of our work in the Appendix.

549 Guidelines:

- 550 • The answer NA means that the paper has no limitation while the answer No
551 means that the paper has limitations, but those are not discussed in the paper.
- 552 • The authors are encouraged to create a separate "Limitations" section in their
553 paper.
- 554 • The paper should point out any strong assumptions and how robust the results
555 are to violations of these assumptions (e.g., independence assumptions, noiseless
556 settings, model well-specification, asymptotic approximations only holding
557 locally). The authors should reflect on how these assumptions might be violated
558 in practice and what the implications would be.
- 559 • The authors should reflect on the scope of the claims made, e.g., if the approach
560 was only tested on a few datasets or with a few runs. In general, empirical
561 results often depend on implicit assumptions, which should be articulated.
- 562 • The authors should reflect on the factors that influence the performance of the
563 approach. For example, a facial recognition algorithm may perform poorly when
564 image resolution is low or images are taken in low lighting. Or a speech-to-text
565 system might not be used reliably to provide closed captions for online lectures
566 because it fails to handle technical jargon.
- 567 • The authors should discuss the computational efficiency of the proposed algo-
568 rithms and how they scale with dataset size.
- 569 • If applicable, the authors should discuss possible limitations of their approach
570 to address problems of privacy and fairness.
- 571 • While the authors might fear that complete honesty about limitations might
572 be used by reviewers as grounds for rejection, a worse outcome might be that
573 reviewers discover limitations that aren't acknowledged in the paper. The
574 authors should use their best judgment and recognize that individual actions in
575 favor of transparency play an important role in developing norms that preserve
576 the integrity of the community. Reviewers will be specifically instructed to not
577 penalize honesty concerning limitations.

578 **3. Theory assumptions and proofs**

579 Question: For each theoretical result, does the paper provide the full set of assump-
580 tions and a complete (and correct) proof?

581 Answer: [NA] .

582 Justification: This work introduces a dataset and benchmark, so the question is not
583 applicable.

584 Guidelines:

- 585 • The answer NA means that the paper does not include theoretical results.
- 586 • All theorems, formulas, and proofs in the paper should be numbered and
587 cross-referenced.
- 588 • All assumptions should be clearly stated or referenced in the statement of any
589 theorems.
- 590 • The proofs can either appear in the main paper or the supplemental material,
591 but if they appear in the supplemental material, the authors are encouraged to
592 provide a short proof sketch to provide intuition.
- 593 • Inversely, any informal proof provided in the core of the paper should be
594 complemented by formal proofs provided in the appendix or supplemental
595 material.
- 596 • Theorems and Lemmas that the proof relies upon should be properly referenced.

597 4. Experimental result reproducibility

598 Question: Does the paper fully disclose all the information needed to reproduce
599 the main experimental results of the paper to the extent that it affects the main
600 claims and/or conclusions of the paper (regardless of whether the code and data are
601 provided or not)?

602 Answer: [Yes]

603 Justification: We share the code to reproduce the results on GitHub and dataset on
604 Kaggle.

605 Guidelines:

- 606 • The answer NA means that the paper does not include experiments.
- 607 • If the paper includes experiments, a No answer to this question will not be
608 perceived well by the reviewers: Making the paper reproducible is important,
609 regardless of whether the code and data are provided or not.
- 610 • If the contribution is a dataset and/or model, the authors should describe the
611 steps taken to make their results reproducible or verifiable.
- 612 • Depending on the contribution, reproducibility can be accomplished in various
613 ways. For example, if the contribution is a novel architecture, describing the
614 architecture fully might suffice, or if the contribution is a specific model and
615 empirical evaluation, it may be necessary to either make it possible for others
616 to replicate the model with the same dataset, or provide access to the model. In
617 general, releasing code and data is often one good way to accomplish this, but
618 reproducibility can also be provided via detailed instructions for how to replicate
619 the results, access to a hosted model (e.g., in the case of a large language model),
620 releasing of a model checkpoint, or other means that are appropriate to the
621 research performed.
- 622 • While NeurIPS does not require releasing code, the conference does require all
623 submissions to provide some reasonable avenue for reproducibility, which may
624 depend on the nature of the contribution. For example
 - 625 (a) If the contribution is primarily a new algorithm, the paper should make it
626 clear how to reproduce that algorithm.
 - 627 (b) If the contribution is primarily a new model architecture, the paper should
628 describe the architecture clearly and fully.
 - 629 (c) If the contribution is a new model (e.g., a large language model), then there
630 should either be a way to access this model for reproducing the results or a
631 way to reproduce the model (e.g., with an open-source dataset or instructions
632 for how to construct the dataset).
 - 633 (d) We recognize that reproducibility may be tricky in some cases, in which
634 case authors are welcome to describe the particular way they provide for

635 reproducibility. In the case of closed-source models, it may be that access to
636 the model is limited in some way (e.g., to registered users), but it should be
637 possible for other researchers to have some path to reproducing or verifying
638 the results.

639 **5. Open access to data and code**

640 Question: Does the paper provide open access to the data and code, with sufficient
641 instructions to faithfully reproduce the main experimental results, as described in
642 supplemental material?

643 Answer: [Yes]

644 Justification: We share the code to reproduce the results on GitHub.

645 Guidelines:

- 646 • The answer NA means that paper does not include experiments requiring code.
- 647 • Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- 648 • While we encourage the release of code and data, we understand that this might
649 not be possible, so “No” is an acceptable answer. Papers cannot be rejected
650 simply for not including code, unless this is central to the contribution (e.g., for
651 a new open-source benchmark).
- 652 • The instructions should contain the exact command and environment needed
653 to run to reproduce the results. See the NeurIPS code and data submis-
654 sion guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>)
655 for more details.
- 656 • The authors should provide instructions on data access and preparation, in-
657 cluding how to access the raw data, preprocessed data, intermediate data, and
658 generated data, etc.
- 659 • The authors should provide scripts to reproduce all experimental results for
660 the new proposed method and baselines. If only a subset of experiments are
661 reproducible, they should state which ones are omitted from the script and why.
- 662 • At submission time, to preserve anonymity, the authors should release
663 anonymized versions (if applicable).
- 664 • Providing as much information as possible in supplemental material (appended
665 to the paper) is recommended, but including URLs to data and code is permitted.

667 **6. Experimental setting/details**

668 Question: Does the paper specify all the training and test details (e.g., data splits,
669 hyperparameters, how they were chosen, type of optimizer, etc.) necessary to
670 understand the results?

671 Answer: [Yes]

672 Justification: All of the training details are in the GitHub code.

673 Guidelines:

- 674 • The answer NA means that the paper does not include experiments.
- 675 • The experimental setting should be presented in the core of the paper to a level
676 of detail that is necessary to appreciate the results and make sense of them.
- 677 • The full details can be provided either with the code, in appendix, or as
678 supplemental material.

679 **7. Experiment statistical significance**

680 Question: Does the paper report error bars suitably and correctly defined or other
681 appropriate information about the statistical significance of the experiments?

682 Answer: [Yes]

683 Justification: Experiment were run using different seeds.

684 Guidelines:

- 685 • The answer NA means that the paper does not include experiments.

- 686 • The authors should answer "Yes" if the results are accompanied by error bars,
 687 confidence intervals, or statistical significance tests, at least for the experiments
 688 that support the main claims of the paper.
- 689 • The factors of variability that the error bars are capturing should be clearly
 690 stated (for example, train/test split, initialization, random drawing of some
 691 parameter, or overall run with given experimental conditions).
- 692 • The method for calculating the error bars should be explained (closed form
 693 formula, call to a library function, bootstrap, etc.)
- 694 • The assumptions made should be given (e.g., Normally distributed errors).
- 695 • It should be clear whether the error bar is the standard deviation or the standard
 696 error of the mean.
- 697 • It is OK to report 1-sigma error bars, but one should state it. The authors
 698 should preferably report a 2-sigma error bar than state that they have a 96%
 699 CI, if the hypothesis of Normality of errors is not verified.
- 700 • For asymmetric distributions, the authors should be careful not to show in
 701 tables or figures symmetric error bars that would yield results that are out of
 702 range (e.g. negative error rates).
- 703 • If error bars are reported in tables or plots, The authors should explain in the
 704 text how they were calculated and reference the corresponding figures or tables
 705 in the text.

706 8. Experiments compute resources

707 Question: For each experiment, does the paper provide sufficient information on the
 708 computer resources (type of compute workers, memory, time of execution) needed
 709 to reproduce the experiments?

710 Answer: [No]

711 Justification: The baseline experiments provided did not require extensive resources.
 712 All experiments can be conducted on a single NVIDIA 3090.

713 Guidelines:

- 714 • The answer NA means that the paper does not include experiments.
- 715 • The paper should indicate the type of compute workers CPU or GPU, internal
 716 cluster, or cloud provider, including relevant memory and storage.
- 717 • The paper should provide the amount of compute required for each of the
 718 individual experimental runs as well as estimate the total compute.
- 719 • The paper should disclose whether the full research project required more
 720 compute than the experiments reported in the paper (e.g., preliminary or failed
 721 experiments that didn't make it into the paper).

722 9. Code of ethics

723 Question: Does the research conducted in the paper conform, in every respect, with
 724 the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

725 Answer: [Yes]

726 Justification: The research meet the NeurIPS Code of Ethics requirements.

727 Guidelines:

- 728 • The answer NA means that the authors have not reviewed the NeurIPS Code
 729 of Ethics.
- 730 • If the authors answer No, they should explain the special circumstances that
 731 require a deviation from the Code of Ethics.
- 732 • The authors should make sure to preserve anonymity (e.g., if there is a special
 733 consideration due to laws or regulations in their jurisdiction).

734 10. Broader impacts

735 Question: Does the paper discuss both potential positive societal impacts and
 736 negative societal impacts of the work performed?

737 Answer: [Yes]

738 Justification: We discuss positive societal impacts in the introduction and potential
739 negative societal impact in the limitation.

740 Guidelines:

- 741 • The answer NA means that there is no societal impact of the work performed.
- 742 • If the authors answer NA or No, they should explain why their work has no
743 societal impact or why the paper does not address societal impact.
- 744 • Examples of negative societal impacts include potential malicious or unintended
745 uses (e.g., disinformation, generating fake profiles, surveillance), fairness consid-
746 erations (e.g., deployment of technologies that could make decisions that unfairly
747 impact specific groups), privacy considerations, and security considerations.
- 748 • The conference expects that many papers will be foundational research and
749 not tied to particular applications, let alone deployments. However, if there
750 is a direct path to any negative applications, the authors should point it out.
751 For example, it is legitimate to point out that an improvement in the quality
752 of generative models could be used to generate deepfakes for disinformation.
753 On the other hand, it is not needed to point out that a generic algorithm for
754 optimizing neural networks could enable people to train models that generate
755 Deepfakes faster.
- 756 • The authors should consider possible harms that could arise when the technology
757 is being used as intended and functioning correctly, harms that could arise when
758 the technology is being used as intended but gives incorrect results, and harms
759 following from (intentional or unintentional) misuse of the technology.
- 760 • If there are negative societal impacts, the authors could also discuss possible
761 mitigation strategies (e.g., gated release of models, providing defenses in addition
762 to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a
763 system learns from feedback over time, improving the efficiency and accessibility
764 of ML).

765 11. Safeguards

766 Question: Does the paper describe safeguards that have been put in place for
767 responsible release of data or models that have a high risk for misuse (e.g., pretrained
768 language models, image generators, or scraped datasets)?

769 Answer: [NA]

770 Justification: All data are sourced from publicly available data source; we release
771 only ZIP-code-level aggregates and remove sensitive attributes, such as race and
772 ethnicity to minimize privacy and discrimination risks.

773 Guidelines:

- 774 • The answer NA means that the paper poses no such risks.
- 775 • Released models that have a high risk for misuse or dual-use should be released
776 with necessary safeguards to allow for controlled use of the model, for example
777 by requiring that users adhere to usage guidelines or restrictions to access the
778 model or implementing safety filters.
- 779 • Datasets that have been scraped from the Internet could pose safety risks. The
780 authors should describe how they avoided releasing unsafe images.
- 781 • We recognize that providing effective safeguards is challenging, and many papers
782 do not require this, but we encourage authors to take this into account and
783 make a best faith effort.

784 12. Licenses for existing assets

785 Question: Are the creators or original owners of assets (e.g., code, data, models),
786 used in the paper, properly credited and are the license and terms of use explicitly
787 mentioned and properly respected?

788 Answer: [Yes]

789 Justification: Every data source, code repository, and pretrained model is formally
790 cited.

791 Guidelines:

- 792 • The answer NA means that the paper does not use existing assets.
- 793 • The authors should cite the original paper that produced the code package or
- 794 dataset.
- 795 • The authors should state which version of the asset is used and, if possible,
- 796 include a URL.
- 797 • The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- 798 • For scraped data from a particular source (e.g., website), the copyright and
- 799 terms of service of that source should be provided.
- 800 • If assets are released, the license, copyright information, and terms of use in
- 801 the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help
- 802 determine the license of a dataset.
- 803 • For existing datasets that are re-packaged, both the original license and the
- 804 license of the derived asset (if it has changed) should be provided.
- 805 • If this information is not available online, the authors are encouraged to reach
- 806 out to the asset's creators.
- 807

808 **13. New assets**

809 Question: Are new assets introduced in the paper well documented and is the
810 documentation provided alongside the assets?

811 Answer: [NA]

812 Justification: We provide well documented for the dataset.

813 Guidelines:

- 814 • The answer NA means that the paper does not release new assets.
- 815 • Researchers should communicate the details of the dataset/code/model as part
- 816 of their submissions via structured templates. This includes details about
- 817 training, license, limitations, etc.
- 818 • The paper should discuss whether and how consent was obtained from people
- 819 whose asset is used.
- 820 • At submission time, remember to anonymize your assets (if applicable). You
- 821 can either create an anonymized URL or include an anonymized zip file.

822 **14. Crowdsourcing and research with human subjects**

823 Question: For crowdsourcing experiments and research with human subjects, does
824 the paper include the full text of instructions given to participants and screenshots,
825 if applicable, as well as details about compensation (if any)?

826 Answer: [NA]

827 Justification: This study involves no crowdsourcing tasks or human-subject experiments;
828 all data were obtained from publicly available sources

829 Guidelines:

- 830 • The answer NA means that the paper does not involve crowdsourcing nor
- 831 research with human subjects.
- 832 • Including this information in the supplemental material is fine, but if the main
- 833 contribution of the paper involves human subjects, then as much detail as
- 834 possible should be included in the main paper.
- 835 • According to the NeurIPS Code of Ethics, workers involved in data collection,
- 836 curation, or other labor should be paid at least the minimum wage in the
- 837 country of the data collector.

838 **15. Institutional review board (IRB) approvals or equivalent for research
839 with human subjects**

840 Question: Does the paper describe potential risks incurred by study participants,
841 whether such risks were disclosed to the subjects, and whether Institutional Review
842 Board (IRB) approvals (or an equivalent approval/review based on the requirements
843 of your country or institution) were obtained?

844 Answer: [NA]

845 Justification: The study does not involve recruiting or interacting with human
846 participants; all data are aggregated from publicly available sources.

847 Guidelines:

- 848 • The answer NA means that the paper does not involve crowdsourcing nor
849 research with human subjects.
- 850 • Depending on the country in which research is conducted, IRB approval (or
851 equivalent) may be required for any human subjects research. If you obtained
852 IRB approval, you should clearly state this in the paper.
- 853 • We recognize that the procedures for this may vary significantly between insti-
854 tutions and locations, and we expect authors to adhere to the NeurIPS Code of
855 Ethics and the guidelines for their institution.
- 856 • For initial submissions, do not include any information that would break
857 anonymity (if applicable), such as the institution conducting the review.

858 16. Declaration of LLM usage

859 Question: Does the paper describe the usage of LLMs if it is an important, original,
860 or non-standard component of the core methods in this research? Note that if
861 the LLM is used only for writing, editing, or formatting purposes and does not
862 impact the core methodology, scientific rigorousness, or originality of the research,
863 declaration is not required.

864 Answer: [NA]

865 Justification: No large language model is used in the dataset construction, bench-
866 marking pipeline, or other core methodology.

867 Guidelines:

- 868 • The answer NA means that the core method development in this research does
869 not involve LLMs as any important, original, or non-standard components.
- 870 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>)
871 for what should or should not be described.