Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Leveraging Multimodal LLMs for Built Environment and Housing Attribute Assessment from Street-View Imagery

Siyuan Yao

syao2@nd.edu

University of Notre Dame

Siavash Ghorbany

University of Notre Dame

Kuangshi Ai

University of Notre Dame

Arnav Cherukuthota

University of Notre Dame

Meghan Forstchen

University of Notre Dame

Alexis Korotasz

University of Notre Dame

Matthew Sisk

University of Notre Dame

Ming Hu

University of Notre Dame

Chaoli Wang

University of Notre Dame

Research Article

Keywords: Built environment evaluation, housing attribute assessment, street-view imagery, computer vision, machine learning, multimodal large language models, visualization dashboard

Posted Date: May 15th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9499711/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Leveraging Multimodal LLMs for Built Environment and Housing Attribute Assessment from Street-View Imagery

Siyuan Yao[1*], Siavash Ghorbany[1], Kuangshi Ai[1*], Arnav Cherukuthota[1], Meghan Forstchen[1], Alexis Korotasz[1], Matthew Sisk[1], Ming Hu[1], Chaoli Wang[1*]

[1*]University of Notre Dame, Notre Dame, IN, 46556, USA.

*Corresponding author(s). E-mail(s): syao2@nd.edu; kai@nd.edu; chaoli.wang@nd.edu;
Contributing authors: sghorban@nd.edu; acheruku@nd.edu; mforstch@nd.edu;
akorotas@nd.edu; msisk1@nd.edu; mhu1@nd.edu;

Abstract

We present a novel framework for automatically evaluating building conditions nationwide in the United States by leveraging large language models (LLMs) and Google Street View (GSV) imagery. By fine-tuning Gemma 3 27B on a modest human-labeled dataset, our approach achieves strong alignment with human mean opinion scores (MOS), outperforming even individual raters on SRCC and PLCC relative to the MOS benchmark. To enhance efficiency, we apply knowledge distillation, transferring the capabilities of Gemma 3 27B to a smaller Gemma 3 4B model that achieves comparable performance with a 3× speedup. Further, we distill the knowledge into a CNN-based model (EfficientNetV2-M) and a transformer (SwinV2-B), delivering close performance while achieving a 30× speed gain. Furthermore, we investigate LLMs' capabilities for assessing an extensive list of built environment and housing attributes through a human–AI alignment study and develop a visualization dashboard that integrates LLM assessment outcomes for downstream analysis by homeowners. Our framework offers a flexible and efficient solution for large-scale building condition assessment, enabling high accuracy with minimal human labeling effort.

Keywords: Built environment evaluation, housing attribute assessment, street-view imagery, computer vision, machine learning, multimodal large language models, visualization dashboard

1 Introduction

The persistent shortage of affordable housing in the United States, aging infrastructure, and rising energy costs burden low-income households [1]. Retrofitting existing housing stock has emerged as a more feasible and cost-effective alternative to new construction, offering opportunities to improve thermal comfort, reduce utility bills, and mitigate health risks from extreme heat and cold [2]. Assessing the physical condition of buildings, especially the exterior envelope (i.e., façades), is critical to determining retrofit needs. However, audits remain resource-intensive, requiring manual inspections and data collection, limiting their practice at a nationwide scale.

Advances in computer vision [3–6] have enabled automated analysis of street-view imagery, such as *Google Street View* (GSV) images, to identify *passive design indicators* (PDIs) like window-to-wall ratios, shading devices, and exterior material types. While these approaches leverage standard computer vision

models (e.g., ResNet) or vision-language models (e.g., BLIP) to classify exterior features with high accuracy, their capacity to simultaneously consider diverse visual elements indicating various features (e.g., paint, window, and structural conditions) and provide interpretable assessments remains limited. *Large language models* (LLMs), especially multimodal LLMs that integrate visual and textual reasoning, present a promising new frontier for interpretable and generalizable assessment. Aligning LLMs with human rater evaluations through fine-tuning on annotated street-view imagery could enable accurate and scalable assessments of building conditions, thereby supporting data-driven passive retrofit strategies across diverse urban contexts.

This paper presents a novel method for fine-tuning multimodal LLMs to automate exterior condition assessments for residential buildings, leading to a pipeline that fine-tunes multimodal LLMs to interpret building exteriors, enabling scalable, automated, and cost-effective assessments. The extension of our prior ISVC 2025 work [7] includes the following. First, we expand the scope of assessment beyond a single housing condition to a more comprehensive set of residential context indicators, enabling richer and more holistic characterization of housing conditions. Second, we conduct a human–AI alignment study to evaluate and improve the consistency between multimodal LLM predictions and expert judgments, thereby enhancing reliability and interpretability. Finally, we develop an interactive visualization dashboard that integrates model outputs into an accessible interface, supporting downstream analysis and decision-making by homeowners and other stakeholders.

2 Related Work

Automated visual assessment of the built environment has become an increasingly active area of research, driven by the growing availability of street-view imagery and advances in machine learning. Traditional computer vision approaches, such as CNNs, have been employed to extract architectural features from images for estimating building condition [8–10]. However, these models often rely on narrow indicators, such as wall paint cracks, and struggle with low accuracy when integrating multiple factors due to limited generalizability.

More recently, vision-language models have demonstrated impressive zero-shot classification performance on visual tasks by leveraging cross-modal alignment between images and text. These models have been applied to identify building components or construction materials [5], as well as to estimate physical walkability [11] and perceived safety [12], directly from street-view imagery without needing extensive labeled data. However, their effectiveness in structured assessment tasks, such as condition scoring guided by formal criteria, remains limited due to insufficient capacity for multi-factor reasoning.

There is a growing interest in using LLMs to extract features of the built environment from street-view images. Several recent studies [13–16] have employed ChatGPT to capture detailed building information, including external features and nearby environmental elements. However, ChatGPT's closed-source nature, lack of support for fine-tuning, and the high cost of API usage at scale limit its practicality for large datasets. Alternatively, fine-tuned open-source LLMs can be trained on labeled datasets for post-earthquake structural damage assessment, successfully performing tasks such as identifying damage severity and classifying affected components [17].

Beyond feature extraction, recent work has explored LLMs as judges to approximate human judgment in subjective, multi-criteria tasks [18, 19], combining scalability with context-aware reasoning and achieving strong agreement with human preferences. However, concerns remain regarding reliability, bias, and prompt sensitivity, highlighting the need for careful validation and human alignment [20, 21]. Recent human-in-the-loop and active learning approaches have further emphasized the importance of incorporating human feedback to improve evaluation quality [22]. Motivated by these findings, we adopt LLMs as judges for large-scale housing attribute assessment and conduct a human-LLM alignment study to ensure consistency with expert criteria.

In this work, we fine-tune open-source LLMs for building condition assessment by leveraging diverse visual features that reflect natural aging and property upkeep. These include indicators like the state of windows, façade paint, and roof materials. In addition, we leverage advanced

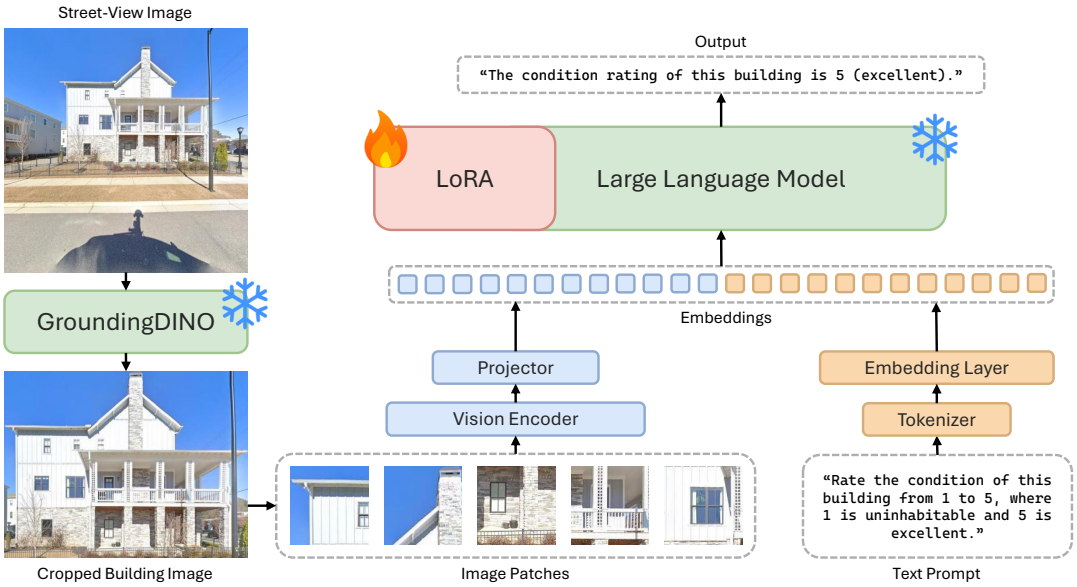


Fig. 1: Overview of our framework for evaluating building conditions.

closed-source multimodal LLMs (e.g., GPT-5.4, Gemini-3.1-Pro, and Claude-Opus-4.6) to extract a broader set of housing attributes beyond overall condition, and align their predictions with human expert annotations through a human–LLM agreement study. Our system produces assessments consistent with expert judgment by aligning model outputs with *mean opinion scores* (MOS) derived from multiple human raters.

3 Our Approach

We aim to evaluate the condition of a building from its GSV image. Experts in architectural research have established formal criteria for condition assessment using a five-point scale (higher ratings indicate better condition). Each rating is defined by detailed descriptors, covering various aspects such as the building's main structure, wall integrity, paint condition, roof, and window quality.

One possible approach is replicating human evaluation by training individual models for each component (e.g., roof, windows, façades), followed by a secondary algorithm to combine these assessments into an overall rating. However, this approach presents several limitations. First, many components may be partially occluded or absent from the image due to angle or visibility. Second, creating labeled datasets for each component would require significant manual effort from experts. Finally, integrating multiple separate models into a single scoring pipeline is complex and inefficient.

To address these challenges, we adopt a more direct strategy: evaluating the entire building holistically using a single model. This method assumes the model can effectively incorporate the full evaluation rubric and infer a corresponding rating. Following the method outlined in [5], and as illustrated in Figure 1, we preprocess images using GroundingDINO to isolate and crop individual buildings from street-view scenes. As found in the experiment in Section 4.2, we choose Gemma 3 [23] as the multimodal LLM backbone to process each cropped image alongside a structured text prompt. In practice, in addition to the query, we include the formal rating criteria in the text prompt as follows:

- **1: Uninhabitable** – Likely unsuitable for rehabilitation; abandoned, fire-damaged, boarded-up, or vacant. Requires demolition.
- **2: Poor** – Requires substantial improvements, including major roof repairs, broken windows, bulging walls, or sagging foundations.

Fig. 2: Examples of buildings corresponding to each condition rating.

- **3: Adequate** – Requires basic cosmetic repairs, with no more than two issues such as painting/-siding, trim, porch, minor roof improvements, or fence repair.
- **4: Good** – Structurally sound with good maintenance and no immediate repairs required. There may be no more than one minor issue, such as limited painting/siding replacement, minor porch repair/painting, or minor fence repair/painting.
- **5: Excellent** – Recently rehabilitated or remodeled; no repairs needed. New paint and roof in very good condition.

Formatting instructions for the expected output are also indicated in the text prompt, enabling the model to evaluate different aspects of the building, including paint, window, structure, and maintenance, according to the criteria, and then provide an overall numerical rating. Figure 2 shows five examples corresponding to ratings 1 to 5, respectively. This overall rating is directly compared with the MOS provided by human experts.

Within this framework, we employ two strategies to leverage Gemma 3: fine-tuning with expert supervision and knowledge distillation for efficiency.

First, to bring the model's predictions closer to the mean opinion of human experts, we fine-tune a strong base model using a small set of human-labeled data. Given the substantial size of modern LLMs and the computational constraints of our local hardware environment, full fine-tuning is not feasible. To address this, we adopt a *parameter-efficient fine-tuning* (PEFT) [24] strategy, which selectively updates a small subset of the model's parameters while keeping the majority fixed. Specifically, we apply *quantized low-rank adaptation* (QLoRA) [25], a technique that enables fine-tuning using low-precision weights and low-rank adapters. This approach significantly reduces memory consumption and training cost, while preserving model performance.

Second, to enable efficient large-scale evaluation, we distill the capabilities of the largest Gemma 3 model (teacher) into smaller, more efficient models (students). For tasks requiring detailed text output, we train a smaller Gemma 3 to replicate the teacher's reasoning and formatting. We transfer knowledge to lightweight vision models for rating-only tasks for faster inference. Distillation is performed using pseudo-labels generated by the teacher on unlabeled building images, removing the need for additional human annotation while preserving alignment with expert judgments.

4 Experiments

4.1 Dataset and Metrics

We collected 12,063 GSV images from six states: California, Florida, Georgia, Indiana, New York, and Texas. The images capture a wide range of building conditions and styles. We invited seven human raters to independently evaluate the building condition of 1,281 randomly selected images on a 1–5 scale. As a result, the labeled images

received MOS ratings of 5 for 229 images, 4 for 582 images, 3 for 345 images, 2 for 105 images, and 1 for 20 images. The remaining 10,782 unlabeled images are used for knowledge distillation experiments. Though resolution varies due to cropping, all images are clear enough for reliable assessment, with a minimum resolution of 600×300.

Since the label distribution is imbalanced (e.g., 1's and 2's are scarce), we emphasize correlation-based evaluation, which is suitable for ordinal MOS labels and robust to skewed class frequencies. We employ the *Spearman's rank correlation coefficient* (SRCC) to evaluate the alignment between model-predicted ratings and human assessments. This non-parametric metric measures the monotonic relationship between two ranked variables, capturing how well the predicted ratings preserve the relative ordering of the ground truth ratings. SRCC is especially well-suited for subjective tasks, where the precise numerical rating may vary across raters, but the relative ranking remains meaningful. In equation,

$$\text{SRCC} = 1 - \frac{6\sum_{i=1}^{N}(x_i - y_i)^2}{N(N^2-1)}, \tag{1}$$

where x_i and y_i denote the model-predicted rating and MOS from human ratings for the i-th image, respectively, and N denotes the number of images. A higher SRCC value indicates stronger agreement in ranking between the model and human raters, with $\rho = 1$ indicating perfect rank correlation.

To assess the linear agreement between model predictions and human ratings, we also compute the *Pearson's linear correlation coefficient* (PLCC). Unlike SRCC, which measures monotonic rank alignment, PLCC quantifies the strength of a linear relationship between predicted ratings and ground truth ratings. It is defined as

$$\text{PLCC} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}, \tag{2}$$

where x_i and y_i denote the model-predicted rating and MOS for the i-th image respectively, and $\bar{x}$, $\bar{y}$ are their corresponding means. N denotes the number of images. A PLCC value of 1 indicates perfect linear correlation, 0 indicates no correlation, and -1 indicates perfect inverse correlation. PLCC is useful when assessing whether the predicted ratings follow the correct ordering and approximate the correct magnitude.

4.2 Zero-shot Evaluation of Multimodal LLMs

We compare the zero-shot performance of multimodal LLMs on building condition evaluation and assess their inference costs. We select the latest open-source models from the LLaVA [26], Mistral [27], LLaMA [28], Qwen [29], and Gemma [23] families. We evaluate two model sizes for LLaVA, Qwen, and Gemma to provide additional reference points. Experiments are run on a system equipped with four NVIDIA A40 GPUs (48 GB VRAM each). While most experiments are performed on a single GPU, models exceeding the memory capacity of one device in our system are run on two GPUs for inference. Note that the multi-GPU execution did not impact inference speed in our experiments. The test set comprises all 1,281 GSV images along with their MOS ratings. All models are given the same text prompt instructing them to evaluate paint, windows, structure, and maintenance, followed by an overall rating from 1 to 5 representing the building condition, which is compared with the MOS to calculate SRCC and PLCC.

As shown in Table 1, LLaMA 4 Scout 17B×16E ('B' denotes the number of parameters in *billions*, while 'E' stands for *experts* in a mixture of experts architecture) achieved the highest correlation with MOS, with an SRCC of 0.78 and a PLCC of 0.79, demonstrating strong agreement with human ratings, which has the same SRCC and PLCC as the average of human raters shown in Table 2. However, because the model requires more than 64 GB of memory, it had to be run on two GPUs in our setup. In comparison, Gemma 3 27B followed closely in performance, achieving a value of 0.77 for SRCC and PLCC. Notably, it requires only 18 GB of VRAM, making it compatible with a single commercial-grade GPU and therefore more practical for deployment on local machines with constrained computational resources. In this comparison, Gemma 3 27B shows the closest alignment with human ratings for building condition evaluation on a single GPU, outperforming larger models such as Mistral Small

Table 1: Comparison of multimodal LLMs. We report SRCC and PLCC with respect to the MOS, along with inference speed (tokens per second), GPU memory usage (VRAM, in gigabytes), and the number of GPUs (nGPU) used to run each model. Inference speed is expressed in tokens per second, ensuring consistency across models regardless of the length of generated text. The best values are highlighted in bold.

model	SRCC ↑	PLCC ↑	inference speed ↑	VRAM ↓	nGPU ↓
Mistral Small 3.2 24B	0.63	0.67	33.86	23.75	**1**
LLaMA 4 Scout 17B×16E	**0.78**	**0.79**	37.28	64.58	2
LLaVA 1.6 7B	0.41	0.41	88.46	5.52	**1**
LLaVA 1.6 34B	0.61	0.61	25.83	25.92	**1**
Qwen 2.5 VL 32B	0.68	0.72	23.66	22.32	**1**
Qwen 2.5 VL 72B	0.73	0.76	11.02	49.77	2
Gemma 3 4B	0.45	0.46	**97.37**	**4.87**	**1**
Gemma 3 27B	0.77	0.77	26.64	18.48	**1**

Table 2: Comparison of individual human raters to the MOS, excluding each rater's own ratings to ensure fairness. We report SRCC and PLCC for each rater, as well as the average across all human raters.

rater	A	B	C	D	E	F	G	average
SRCC ↑	0.78	0.80	0.80	0.76	0.76	0.74	0.81	0.78
PLCC ↑	0.80	0.80	0.81	0.78	0.77	0.76	0.82	0.79

3.2 24B, LLaVA 1.6 34B, Qwen 2.5 VL 32B, and Qwen 2.5 VL 72B.

Performance consistently declines among LLaVA, Qwen, and Gemma as model size is reduced. For Qwen 2.5 VL, the drop in both SRCC and PLCC from 72B to 32B is less than 0.1, which is much smaller than the 0.2 drop observed from 34B to 7B for LLaVA 1.6 or the drop of more than 0.3 from 27B to 4B for Gemma 3. As a result, although Gemma 3 4B demonstrated the highest efficiency in inference speed and the lowest VRAM usage, its predicted ratings deviate substantially from human ratings, making it unsuitable for use.

Overall, Gemma 3 27B is the most practical out-of-the-box open-source choice for building condition evaluation on a single GPU, combining near-human MOS alignment with moderate speed and GPU memory requirements.

4.3 Optimization of Gemma 3

Building on the findings in Section 4.2, we investigate flexible ways to leverage Gemma 3 for building condition evaluation. We first examine whether different output formats in the prompt affect the performance of the Gemma 3 27B model. As shown in Table 3, we experiment with alternative formats for the overall rating, including using a word instead of a number, and restricting the model to output only a single number or word without additional text. This latter case is motivated by the fact that the number of tokens in the response directly impacts inference time. All experiments are evaluated on the same set of 1,281 images. The results in Table 3 indicate that when detailed descriptions are provided, the model's accuracy remains almost consistent regardless of whether the overall rating is expressed as a word or a number. When only the rating is needed, using a single word yields nearly the same accuracy but with much faster responses due to fewer generated tokens. However, accuracy drops significantly when restricted to outputting only a single number. We attribute this phenomenon to numbers with less semantic context than descriptive words, making it harder for the model to link to specific building conditions when used alone.

Next, we fine-tune Gemma 3 27B using MOS labels to improve its alignment with human ratings. Since the accuracy of the word-only response is comparable to that of detailed responses (see Table 3), we adopt the word-only format for these experiments to simplify loss computation based on MOS, which is translated from numerical ratings into their corresponding descriptive words. From the dataset of 1,281 images, the first 800 images

Table 3: Comparison of different output formats. We report SRCC and PLCC, along with the average number of tokens in the response generated, the average response generation time (in seconds), and the prompt processing time per image (in seconds). The best values are highlighted in bold.

output	details & number	details & word	single number	single word
SRCC ↑	0.77	0.77	0.70	**0.78**
PLCC ↑	0.77	**0.78**	0.66	**0.78**
# response tokens ↓	87.82	79.89	**2.00**	**2.00**
response time ↓	3.29	3.00	**0.08**	**0.08**
processing time ↓	1.06	1.04	**0.96**	1.13

Table 4: Comparison of different numbers of training images used to fine-tune the Gemma 3 27B model. We report SRCC and PLCC, along with the training time (in minutes). The best values are highlighted in bold.

# training images	0	100	200	300	400	500	600	700	800
SRCC ↑	0.78	0.77	0.78	0.80	0.81	0.82	0.82	**0.83**	**0.83**
PLCC ↑	0.79	0.78	0.79	0.81	0.81	**0.83**	0.81	0.82	0.82
training time ↓	—	**4.52**	8.20	12.68	16.57	21.48	25.38	30.52	34.68

are allocated for training and the remaining 481 for testing. As shown in Table 4, we vary the number of training images by randomly sampling subsets from the 800 training images. The slightly higher PLCC for the pre-trained Gemma 3 27B model compared to earlier results is attributable to the change in the test set. During fine-tuning, the model is quantized to 4-bit precision to reduce memory usage. Following the Gemma 3 official guidelines, the LoRA configuration uses a scaling factor of 16 and a dropout rate of 0.05 to prevent overfitting. The rank of the low-rank matrix is set to 16, and the warm-up ratio is set to 0.03 to improve training stability. LoRA limits training to roughly 16% of the model's parameters. The learning rate is set to 5×10^{-5}. The training batch size is 1, and only one epoch is run to avoid overfitting. The model is optimized with a next-token prediction loss, calculated as the cross-entropy between the predicted logits and the target labels.

Table 4 shows that fine-tuning Gemma 3 27B on 500 labeled images yields performance that surpasses the MOS alignment of all individual human raters (see Table 2). Increasing the number of training images beyond 500 does not yield definitive performance improvements, indicating that the model reaches a performance plateau. These results suggest that a fine-tuned Gemma 3 27B model trained on 500 images is sufficient to match or exceed human-level consistency, making it a practical replacement for manual rating in automated building condition evaluation.

4.4 Knowledge Distillation using Fine-tuned Gemma 3

We also experiment with response-based knowledge distillation using our Gemma 3 27B fine-tuned on 500 labeled images as a teacher model to fine-tune the Gemma 3 4B model, as well as several efficient vision models, including ResNet [30], MobileNetV3 [31], EfficientNetV2 [32], and Swin Transformer V2 [33]. For comparison, we prepare two training sets: one containing 10,782 building images labeled automatically by the fine-tuned Gemma 3 27B, and another containing 800 human-labeled images. Both are evaluated using the same 481 images. For the Gemma 3 4B model, we apply the same QLoRA fine-tuning procedure described in Section 4.3. For the vision models, we optimize using *mean squared error* (MSE) loss between the predicted ratings and the MOS, with a learning rate of 1×10^{-4} during fine-tuning. The vision models are trained for up to 10 epochs on the Gemma 3–labeled dataset and up to 100 epochs on the human-labeled dataset. We identify the epoch that achieves the highest SRCC on the test set and report its performance.

As shown in Table 5, knowledge distillation effectively transfers the capabilities of the fine-tuned Gemma 3 27B model to smaller models. For all models, fine-tuning on the automatically labeled dataset leads to consistently higher performance than fine-tuning on the human-labeled

Table 5: Comparison of models fine-tuned on two training datasets labeled respectively by Gemma 3 27B and by human annotators. We report SRCC and PLCC for both datasets, along with the batch size used during training, inference speed (in images per second), and GPU memory usage at the inference stage (VRAM, in gigabytes). Inference speed is expressed in images per second because models other than Gemma 3 produce numeric outputs instead of text. The best values are highlighted in bold.

model	Gemma dataset		human dataset		batch	inference	
	SRCC ↑	PLCC ↑	SRCC ↑	PLCC ↑	size	speed ↑	VRAM ↓
ResNet-50	0.68	0.66	0.52	0.53	32	69.41	1.87
MobileNetV3-L	0.65	0.66	0.45	0.46	32	**69.91**	**1.72**
EfficientNetV2-M	0.73	0.74	0.60	0.61	16	31.01	2.09
SwinV2-B	0.73	0.74	0.61	0.63	16	32.61	2.36
Gemma 3 4B	**0.81**	**0.80**	**0.74**	**0.73**	1	3.05	9.10

dataset. The approach performs well across different architectures, including CNNs (ResNet-50, MobileNetV3-L, EfficientNetV2-M), transformers (SwinV2-B), and multimodal LLMs (Gemma 3 4B). This result indicates that the fine-tuned Gemma 3 27B model can serve as a dependable substitute for human annotators in creating a large-scale dataset for building condition evaluation, helping to overcome the scarcity of annotated data while significantly reducing labeling costs. Notably, EfficientNetV2-M and SwinV2-B achieve SRCC and PLCC values above 0.7, comparable to the base Gemma 3 27B model, while delivering more than 30× faster inference. Fine-tuning Gemma 3 4B on the automatically labeled dataset enables it to achieve SRCC and PLCC values exceeding the base Gemma 3 27B model, with the added benefit of roughly 3× faster inference. In this experiment, the fine-tuned LoRA adapters were not merged into the pre-trained model prior to inference. Instead, they were loaded separately at runtime, allowing us to keep only the compact LoRA files rather than full model checkpoints. This approach increased VRAM usage from less than 5 GB to over 9 GB, as both the base model and the adapter were held in memory during inference. Still, users can merge the adapters into the base model, reducing the VRAM consumption to a level close to the pre-trained model. Leveraging knowledge distillation from the fine-tuned Gemma 3 27B model enables scalable, fully automated evaluation with the flexibility to select models for different performance–efficiency trade-offs, making it practical to process datasets containing millions of images within reasonable timeframes.

5 Extended Attribute Extraction

In addition to predicting overall building condition, we examine whether multimodal LLMs can extract a broader set of housing attributes from street-view imagery. These attributes are intended to enrich the downstream visualization dashboard by describing not only the target house itself but also aspects of its visible surroundings. We formulate all attribute extraction tasks as multiple-choice question answering (QA) so that model outputs and human judgments are directly comparable. This setup allows us to evaluate three complementary properties: the stability of repeated LLM assessments, the inter-rater reliability of human experts and LLMs, and the alignment between LLM predictions and human expert assessments.

5.1 Attribute Extraction with Multimodal LLMs

Unlike the experiments in Section 3, which use cropped building views, this study uses zoomed-out street-view images that preserve the focal building together with its immediate context. The goal is to determine whether multimodal LLMs can infer attributes that depend not only on the house itself, such as structural condition and architectural features, but also on visible neighborhood cues, including walkability, safety, accessibility, and social environment, as well as environmental context, including landscaping condition, environmental hazards, and geographic setting. We collected 143 additional images spanning diverse geographic settings and housing conditions. Each image contains a single primary building while

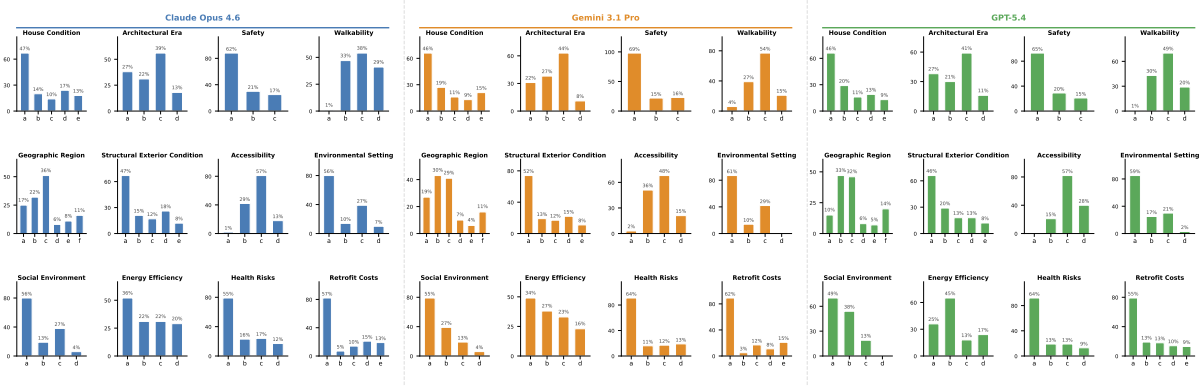


Fig. 3: Attribute-label distributions for GPT-5.4, Gemini-3.1-Pro, and Claude-Opus-4.6 after majority voting across five runs.

retaining enough surrounding context to support richer attribute assessment. For every image, we consider 12 attributes in total: the main housing-condition attribute and 11 additional attributes defined by architecture experts. The complete attribute definitions and prompting protocol are provided in Appendix A.

We evaluate three proprietary multimodal LLMs: GPT-5.4, Gemini-3.1-Pro, and Claude-Opus-4.6. To ensure a fair comparison across models, we constrain each model to return exactly one categorical label for each attribute rather than an open-ended explanation. In other words, the task is cast as a multiple-choice QA problem. At inference time, the model receives the building image together with the 12 attribute questions and their candidate labels, and it must output one label per attribute. To reduce possible order effects, we randomly shuffle the order of the attributes for every run. Each model is queried five times per image, which enables us to evaluate both prediction quality and run-to-run consistency. Figure 3 shows the resulting attribute-label distributions for the three LLM judges.

5.2 Human Expert Annotations

To assess how closely LLM judgments align with expert opinion, we recruited six annotators with architecture expertise and asked them to complete the same multiple-choice task. Because this annotation process is substantially more time-intensive for humans than for models, we selected a representative subset of 25 images covering diverse geographic regions, neighborhood contexts, and levels of housing condition. The experts were shown the same zoomed-out images and the same attribute definitions used in model evaluation, and they assigned a single categorical label to each attribute for each image. This expert-labeled subset contains 300 image-attribute pairs and serves two purposes: it allows us to measure human inter-rater reliability and provides a reference for evaluating human-LLM alignment.

5.3 Experimental Results

We first evaluate whether each LLM is internally consistent across repeated trials with different random attribute orders. We define the *stability score* as the average pairwise agreement across repeated runs. Specifically, for each image-attribute pair, we compare the model outputs across all $\binom{5}{2} = 10$ run pairs and compute the fraction of run pairs that produce the same categorical label. We then average this fraction over all image-attribute pairs. A higher stability score, therefore, indicates that the model is less sensitive to prompt-order variation. As auxiliary dispersion summaries, for each image-attribute pair, we compute the standard deviation of the five repeated assessments on the ordered label scale, and then report the mean of these per-pair standard deviations over all image-attribute pairs. Lower values indicate more stable predictions. The results in Table 6 reveal differences in model robustness to prompt-order variation. Claude-Opus-4.6 achieves the highest stability score and lowest variance, which aligns with its more concentrated distributions observed

in Figure 3. In contrast, GPT-5.4 exhibits the lowest stability and highest dispersion, and Gemini-3.1-Pro provides a balanced trade-off.

Table 6: Stability of LLM assessments across 5 repeated trials with random attribute orders. A higher stability score indicates greater invariance to attribute ordering, while a lower mean standard deviation indicates less dispersion across runs. The best values are highlighted in bold.

model	stability score ↑	avg std ↓
Gemini-3.1-Pro	0.896	0.096
Claude-Opus-4.6	**0.941**	**0.054**
GPT-5.4	0.861	0.131

Next, we compare the internal agreement of human experts and LLM judges using Krippendorff's α and the intraclass correlation coefficient, ICC(2,1). Krippendorff's α measures the extent to which raters assign consistent categorical labels, while ICC summarizes agreement when judgments are treated as ordered ratings. For both metrics, higher values indicate stronger inter-rater reliability. Table 7 reports the agreement results for human experts, individual models, and the cross-model aggregate. Overall, all LLMs achieve substantially higher inter-rater reliability than human experts. Specifically, Krippendorff's α and ICC values exceed 0.94 for each model and remain above 0.92 when aggregating across different models, indicating strong consistency both within and between LLMs. This suggests that LLM judgments are far more internally consistent than human annotations. However, higher agreement does not necessarily imply higher correctness. Instead, it reflects the more deterministic and less subjective nature of LLM outputs compared to human reasoning. Notably, Claude-Opus-4.6 achieves the highest agreement, reinforcing its strong stability observed in Table 6.

Finally, we evaluate how well LLM judgments align with human expert evaluations on the expert-annotated subset. For each image-attribute pair, human ratings are determined by majority voting across experts, and model predictions are determined by majority voting across repeated runs. We then compute Pearson correlation, Spearman rank correlation, *mean absolute error* (MAE), and *root mean square error* (RMSE) between the aggregated human and model judgments. Together, these metrics capture both rank-order consistency and absolute deviation from expert assessments.

Table 7: Inter-rater reliability within human experts and LLMs (Gemini-3.1-Pro, Claude-Opus-4.6, GPT-5.4), measured by Krippendorff's α and ICC(2,1). Here, n denotes the number of completed assessments, where each assessment is a full annotation over all 1,716 image-attribute pairs by one expert or one model run. The best values are highlighted in bold.

evaluator	Krippendorff's α ↑	ICC(2,1) ↑
Human Experts (n=6)	0.635	0.640
Gemini-3.1-Pro (n=5)	0.956	0.960
Claude-Opus-4.6 (n=5)	**0.979**	**0.980**
GPT-5.4 (n=5)	0.940	0.940
Cross LLMs (n=15)	0.920	0.918

As shown in Table 8, all three models demonstrate strong alignment with human expert judgments, with correlation values above 0.78. Gemini-3.1-Pro achieves the best overall performance across both correlation and error metrics, indicating that it most closely matches human assessments despite not having the highest internal consistency. This suggests that higher stability or inter-model agreement does not necessarily translate to better alignment with human judgments. In contrast, Claude-Opus-4.6 shows slightly lower agreement with human annotations, indicating a possible bias toward more deterministic but less human-aligned predictions.

Overall, the results from Figure 3, Table 6, Table 7, and Table 8 reveal a consistent pattern: models that produce more concentrated label distributions tend to achieve higher stability and inter-rater reliability, but not necessarily better alignment with human judgments. This suggests that different models occupy distinct points along a spectrum between consistency and human fidelity, providing flexibility for downstream applications depending on whether robustness or human alignment is prioritized.

Table 8: Alignment between LLM assessments and human expert evaluations on the expert-annotated subset (25 images; 300 image-attribute pairs). Higher correlation and lower error indicate closer agreement with human assessments. The best values are highlighted in bold.

model	Pearson r ↑	Spearman ρ ↑	MAE ↓	RMSE ↓
Gemini-3.1-Pro	**0.844**	**0.842**	**0.327**	**0.616**
Claude-Opus-4.6	0.803	0.800	0.423	0.705
GPT-5.4	0.788	0.798	0.390	0.705

6 Visualization Dashboard

While Sections 3 to 5 focus on model development and evaluation, practical deployment also requires a user-facing interface to present these outputs in a form suitable for inspection and downstream decision-making. This motivates us to develop a visualization dashboard that connects the image-based building assessment with a broader building-analysis workflow. Figure 4 shows a screenshot of the dashboard. At a high level, the dashboard integrates property information, map-based building lookup, LLM-assisted assessment results, and city-level comparison views in a single interface. The left panel supports building-level review and retrofit-oriented exploration, the central map provides parcel-level spatial context, and the right panel places the selected building within broader municipal patterns. In this paper, we include the dashboard to illustrate one possible way the proposed assessment pipeline can be presented to end users (e.g., homeowners), without attempting a full evaluation of the interface itself.

Among the dashboard components, the AI analysis views shown in Figure 5a and Figure 5b are the ones most closely connected to this work. Building on the additional attribute extraction discussed in Section 5, the dashboard combines street-view imagery with structured building attributes and neighborhood context, and then uses an LLM-based module to produce category-level assessments such as structural integrity, safety, accessibility, environmental quality, social environment, and energy efficiency. The interface is intended to make these outputs easier to inspect by presenting compact scores, confidence cues, and supporting details in a format that is more actionable than a standalone model prediction.

The report view in Figure 5b complements the summary scores with a more explicit audit trail. Instead of returning only a single score, the system organizes basic property attributes, observed exterior conditions, and short textual rationales into a downloadable narrative report. This design is useful for practical review because it preserves the connection between the model output and the evidence available from street-view imagery and auxiliary records. At the same time, we keep the discussion here intentionally limited: components such as interactive chat, retrofit exploration, financing guidance, and broader city-level analytics are part of the dashboard environment, but a full evaluation of those modules is beyond the scope of this paper.

7 Conclusions and Future Work

We present a novel framework for automated building condition evaluation from GSV images across the United States using multimodal LLMs with minimal human annotation. We benchmark multiple leading open-source multimodal LLMs, identify the most effective model for aligning with expert ratings, and explore techniques, such as prompt refinement, targeted fine-tuning, and knowledge distillation, to enhance reliability and significantly improve efficiency, making million-scale dataset evaluation practical. The framework provides a flexible selection of methods, allowing users to balance performance and efficiency according to task-specific priorities. We further examine the capabilities of LLMs to assess an extensive list of built environment and housing attributes through a human–AI alignment study. Moreover, we develop a visualization dashboard that integrates LLM assessment outcomes for downstream analysis by homeowners. While our work is the first to evaluate building quality at a large scale using LLMs, a promising future direction is to explore their application to broader aspects of buildings.

Still, our methods have a few limitations. First, the MOS ratings are derived from a specific group of skilled and novice raters, which may not fully capture broader subjective consensus or

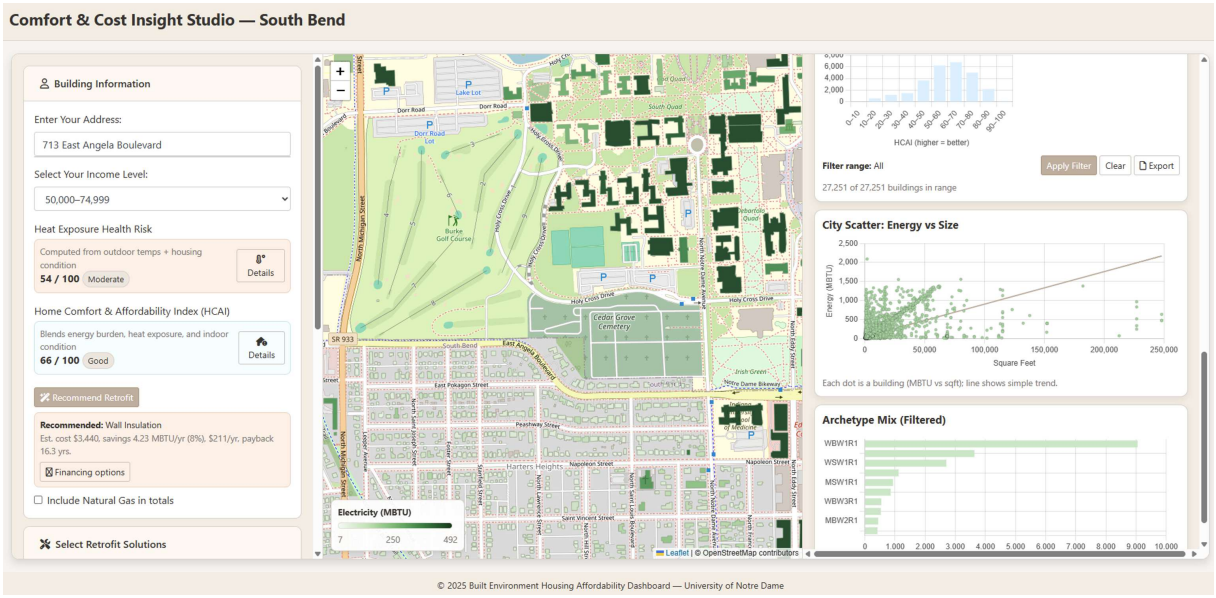
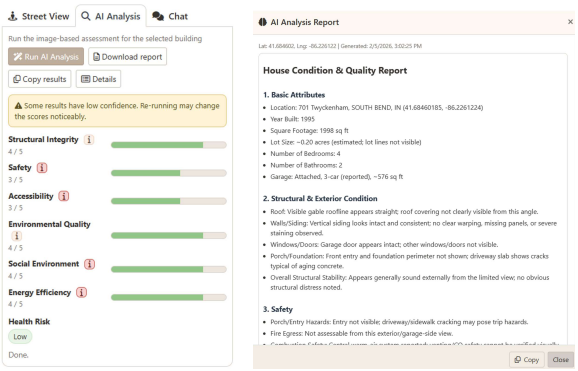


Fig. 4: Overview of the visualization dashboard, which integrates building-level property review, parcel-level map context, and city-scale comparison views in a unified interface.



(a) analysis tabs (b) LLM report

Fig. 5: Example dashboard views for LLM-based building assessment: the analysis tabs provide a compact summary of model outputs, while the report view presents detailed evaluation results for the selected property.

ensure complete fairness. While this may be sufficient for meeting the needs of a small target group, broader application scenarios require input from a larger and more diverse rater pool. Second, output-based knowledge distillation demands large quantities of raw images to generate pseudo-labeled image–rating pairs from real-world data. While feature-based knowledge distillation may offer greater efficiency, differences in network architectures introduce challenges in maintaining compatibility between the teacher and student feature representations. We consider exploring this as a future research direction. Third, given that LLM performance is highly dependent on prompt construction, future work will explore more sophisticated prompt engineering strategies beyond the plain-language criteria, query, and format descriptions used in this study. Finally, future work could examine potential biases in LLM outputs stemming from various image attributes, both low-level factors such as brightness and contrast, and high-level characteristics such as architectural style or building type, to better understand influences on model predictions.

In addition, future work will expand the visualization dashboard into a more thoroughly evaluated decision-support system. In particular, we plan to study how the AI analysis reports, additional attribute extraction, retrofit recommendation workflow, and city-scale visualization modules can be integrated and validated for different user groups, including residents, researchers, and local agencies.

Acknowledgements. This research was supported in part by the U.S. National Science

Foundation through grants IIS-1955395, IIS-2101696, OAC-2104158, TI-2317971, IIS-2401144, and CNS-2430623, the University of Notre Dame's Just Transformations to Sustainability (JTS) and Data, AI, and Computing (DAC) Initiatives, Lucy Family Institute for Data & Society Health Equity Data Lab, and Notre Dame-IBM Technology Ethics Lab.

Appendix A Extended Attribute Extraction

To examine whether advanced closed-source multimodal LLMs can support broader housing analysis beyond the primary house-condition rating studied in the main paper, we prepared an additional prompt specification covering 12 attributes: one core housing-condition attribute and 11 additional contextual attributes. In practice, each model receives one street-view image together with the attribute definitions below and is asked to return exactly one label per attribute.

A.1 Generic Evaluation Prompt

The following template summarizes the prompt structure used for the three advanced closed-source models.

Role. You are an expert assessor of residential buildings from street-view imagery.

Task. Given a single exterior image of a house, evaluate the property using only visible evidence. Do not infer details that cannot be reasonably observed.

Instructions.

- For each attribute below, choose exactly one label from the provided options.
- Base each judgment only on what is visible in the image.
- If evidence is limited, select the most conservative plausible label.
- Keep the output concise and structured.

Required output format.

- House Condition: <label>
- Architectural Era: <label>
- Safety: <label>
- Walkability: <label>
- Geographic Region: <label>
- Structural & Exterior Condition: <label>
- Accessibility: <label>
- Environmental Setting: <label>
- Social Environment: <label>
- Energy Efficiency: <label>
- Health Risks: <label>
- Retrofit & Costs: <label>

Use the attribute definitions exactly as provided.

A.2 Attribute Definitions

A.2.1 House Condition

- **Excellent:** Recently rehabilitated or remodeled; no repairs needed. New paint and roof in very good condition.
- **Good:** Structurally sound with good maintenance and no immediate repairs required. At most one minor issue is visible, such as limited painting or siding replacement, minor porch repair or painting, or minor fence repair or painting.
- **Adequate:** Requires basic cosmetic repairs, with no more than two issues such as painting or siding work, trim repair, porch repair, minor roof improvement, or fence repair.
- **Poor:** Requires substantial improvements, including major roof repairs, broken windows, bulging walls, or sagging foundations.
- **Uninhabitable:** Likely unsuitable for rehabilitation; abandoned, fire-damaged, boarded-up, or vacant, and likely requires demolition.

A.2.2 Architectural Era (Estimated Year Built)

- **Pre-1950 (Historic):** Hand-laid brick or stone rather than veneer, true plaster, and wood windows or doors. Window openings tend to be smaller, with wood double-hung sash and divided panes. Taller-than-wide windows, visible chimneys, and crafted wood trim or porch elements are common.
- **1950–1980 (Mid-Century):** Long, low one-story ranches or split-level forms. Large single-pane picture windows with aluminum or steel frames, often in horizontal groupings. Minimal trim, low-pitch roofs, simple facades, and emphasized garages or carports are common.
- **1981–2010 (Modern Suburban):** Two-story massing with complex rooflines and multiple

gables. Mixed-material facades such as brick or stone veneer combined with siding. Vinyl double-hung or casement windows, decorative grilles, arched shapes, and prominent attached garages are common.
- **2011–Present (Contemporary):** Simplified or boxy forms with cleaner lines, modern flat or low-slope roofs or refined suburban styles, fiber-cement or engineered siding, manufactured stone, high-contrast color palettes, and larger energy-efficient windows, often with dark frames.

A.2.3 Safety

- **Secure:** Safe entryways, intact railings, and stable structural lines.
- **Compromised:** Minor hazards such as cracked steps, missing brick, or aging railings.
- **Hazardous:** Visible structural failures, sagging roofs, leaning columns, broken foundations, or blocked exits.

A.2.4 Walkability

- **Pedestrian-Integrated:** Clear paved sidewalks on both sides of the street, with visible crosswalks or bike lanes.
- **Limited Connectivity:** Sidewalks are present but disconnected or only on one side of the street; the area remains primarily car-oriented.
- **Car-Dependent:** No sidewalks or footpaths are visible; properties transition directly into the street or a large parking lot.
- **Primitive / Rural:** Unpaved roads with no visible public infrastructure for pedestrians or transit.

A.2.5 Geographic Region

- **Northeast**
- **Midwest**
- **Southeast / Gulf Coast**
- **Southwest / Desert (AZ, NM, NV)**
- **West Coast / Pacific Northwest**
- **Mountain / Northern Rockies (CO, UT, MT, ID, WY)**

A.2.6 Structural & Exterior Condition

- **Excellent:** Roof, siding, windows, and foundation all appear new or recently maintained, with no visible damage or wear.
- **Good:** Minor surface wear is visible on one element, such as slight paint fading or light weathering, with no structural concerns.
- **Fair:** Moderate deterioration is visible on one or two elements, such as an aging roof, worn siding, or hairline foundation cracks.
- **Poor:** Significant damage is visible across multiple elements, such as a sagging roof, broken windows, crumbling walls, or foundation issues.
- **Critical:** Severe structural-failure risk is visible, with near-collapse condition, extensive exterior damage, or major foundation failure.

A.2.7 Accessibility

- **Fully Accessible:** Step-free entry, wide doorways, and a level approach are clearly visible.
- **Mostly Accessible:** One or two minor barriers are visible, such as a single step at the entry, but the property is otherwise manageable.
- **Limited Accessibility:** Several barriers are visible, such as multiple steps, a narrow entry, or no ramp or handrail.
- **Not Accessible:** Significant barriers are visible throughout, with no clear mobility accommodations.

A.2.8 Environmental Setting

- **Well-Maintained:** The yard, landscaping, and surroundings appear neat and clean, with no visible environmental hazards.
- **Average:** Some yard maintenance is needed, but the surroundings are typical and show no obvious hazards.
- **Neglected:** The yard or surroundings appear overgrown, cluttered, or poorly maintained, with minor environmental concerns.
- **High Risk:** Visible environmental hazards are present, such as flooding indicators, industrial proximity, or a severely degraded lot or surrounding area.

A.2.9 Social Environment

- **Strong:** Neighboring properties are well maintained, and clean streets or active landscaping suggest strong community investment.
- **Stable:** Nearby properties show mixed maintenance levels, but the neighborhood fabric appears generally intact.
- **Declining:** Several neglected or vacant neighboring properties are visible, with limited visible community investment.
- **Distressed:** Surrounding properties are predominantly vacant, boarded-up, or heavily deteriorated.

A.2.10 Energy Efficiency

- **High Efficiency:** Newer construction or visible retrofit signs such as insulated siding, double- or triple-pane windows, or modern HVAC equipment are present.
- **Moderate Efficiency:** A mix of older and newer features is visible, with standard windows and typical systems for the era.
- **Low Efficiency:** Older single-pane windows, limited visible weatherization, or aging systems are apparent.
- **Very Low Efficiency:** The home appears pre-1950 with no visible upgrades and no evidence of weatherization or insulation improvements.

A.2.11 Health Risks

- **Low Risk:** Surfaces are well maintained, with no visible mold, peeling paint, or moisture damage; the home may also be newer.
- **Moderate Risk:** Minor peeling paint or light moisture staining is visible, especially for older homes with mostly intact surfaces.
- **High Risk:** Visible peeling paint, potential mold or moisture staining, or significant deterioration is present on an older home.
- **Severe Risk:** Extensive deterioration is visible together with multiple health hazards, such as abandonment, fire damage, or partial collapse.

A.2.12 Retrofit & Costs

- **Move-in Ready:** No immediate retrofit is needed beyond cosmetic updates.
- **Light Retrofit:** Minor repairs are needed on one or two elements, implying low investment.
- **Moderate Retrofit:** Several components or systems visibly require replacement, implying moderate investment.
- **Heavy Retrofit:** Major structural, mechanical, or envelope work is clearly needed, implying significant investment.
- **Full Rehabilitation:** Near-complete renovation is required, corresponding to the highest investment level.

References

[1] Anacker, K.B.: Introduction: Housing affordability and affordable housing. International Journal of Housing Policy **19**(1), 1–16 (2019)

[2] Hu, M., Ghorbany, S., Yao, S., Wang, C., Sisk, M.: BUILT2AFFORD: Machine-learning-driven passive retrofits for affordable housing. In: Proceedings of Architectural Research Centers Consortium Annual Conference (2025)

[3] Starzyńska-Grześ, M.B., Roussel, R., Jacoby, S., Asadipour, A.: Computer vision-based analysis of buildings and built environments: A systematic review of current approaches. ACM Computing Survey **55**(13s), 284–128425 (2023)

[4] Ghorbany, S., Hu, M., Yao, S., Wang, C., Nguyen, Q.C., Yue, X., Alirezaei, M., Tasdizen, T., Sisk, M.: Examining the role of passive design indicators in energy burden reduction: Insights from a machine learning and deep learning approach. Building and Environment **250**, 111126 (2024)

[5] Yao, S., Ghorbany, S., Sisk, M., Hu, M., Wang, C.: Leveraging zero-shot learning on street-view imagery for built environment variable analysis. In: Proceedings of International Symposium on Visual Computing, pp. 243–254 (2024)

[6] Ghorbany, S., Hu, M., Yao, S., Sisk, M., Wang, C., Zhang, K., Nguyen, Q.C.: Data driven assessment of built environment impacts on urban health across United States cities. Scientific Reports **15**, 19998 (2025)

[7] Yao, S., Ghorbany, S., Forstchen, M., Korotaszand, A., Sisk, M., Hu, M., Wang, C.: Leveraging multimodal LLMs for building condition assessment from street-view imagery. In: Proceedings of International Symposium on Visual Computing, pp. 219–231 (2025)

[8] Hoang, N.-D.: Image processing-based recognition of wall defects using machine learning approaches and steerable filters. Computational Intelligence and Neuroscience **2018**, 7913952 (2018)

[9] Amrouni Hosseini, M., Ravanshadnia, M., Rahimzadegan, M., Ramezani, S.: Next-generation building condition assessment: BIM and neural network integration. Journal of Performance of Constructed Facilities **38**(6), 04024050 (2024)

[10] Zou, S., Wang, L.: Detecting individual abandoned houses from Google street view: A hierarchical deep learning approach. ISPRS Journal of Photogrammetry and Remote Sensing **175**, 298–310 (2021)

[11] Liu, X., Haworth, J., Wang, M.: A new approach to assessing perceived walkability: Combining street view imagery with multimodal contrastive learning model. In: Proceedings of ACM SIGSPATIAL International Workshop on Spatial Big Data and AI for Industrial Applications, pp. 16–21 (2023)

[12] Wang, X., Gilvear, A., Li, Y., Ilyankou, I.: Can CLIP see safe streets? comparing human and VLM perceptions of walkability and safety. In: Proceedings of AGILE Walking the X-min City Workshop (2025)

[13] Cheng, Y., Yin, Z., Li, D., Li, Z.: Assessing urban safety: A digital twin approach using streetview and large language models. In: Proceedings of IEEE Vehicular Technology Conference, pp. 1–5 (2024)

[14] Li, Z., Su, Y., Wang, H., Zhao, W.: BuildingView: Constructing urban building exteriors databases with street view imagery and multimodal large language mode. arXiv preprint arXiv:2409.19527 (2024)

[15] Malekzadeh, M., Willberg, E., Torkko, J., Toivonen, T.: Urban attractiveness according to ChatGPT: Contrasting AI and human insights. Computers, Environment and Urban Systems **117**, 102243 (2025)

[16] Liang, X., Xie, J., Zhao, T., Stouffs, R., Biljecki, F.: OpenFACADES: An open framework for architectural caption and attribute data enrichment via street view imagery. arXiv preprint arXiv:2504.02866 (2025)

[17] Jiang, Y., Wang, J., Shen, X., Dai, K.: Large language model for post-earthquake structural damage assessment of buildings. Computer-Aided Civil and Infrastructure Engineering (2025)

[18] Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., et al.: A survey on LLM-as-a-judge. The Innovation (2026). In Press

[19] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., *et al.*: Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In: Proceedings of Annual Conference on Neural Information Processing Systems, pp. 46595–46623 (2023)

[20] Shankar, S., Zamfirescu-Pereira, J.D., Hartmann, B., Parameswaran, A., Arawjo, I.: Who validates the validators? aligning LLM-assisted evaluation of LLM outputs with human preferences. In: Proceedings of ACM Symposium on User Interface Software and Technology, pp. 131–113114 (2024)

[21] Chehbouni, K., Haddou, M., Cheung, J.C.K., Farnadi, G.: Neither valid nor reliable? investigating the use of LLMs as judges. arXiv preprint arXiv:2508.18076 (2025)

[22] Gebreegziabher, S.A., Ai, K., Zhang, Z., Glassman, E., Li, T.J.-J.: Leveraging variation theory in counterfactual data augmentation for optimized active learning. In: Proceedings of Findings of the Association for Computational Linguistics, pp. 894–906 (2025)

[23] Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M.S., Love, J., Tafti, P., et al.: Gemma: Open models based on Gemini research and technology. arXiv preprint arXiv:2403.08295 (2024)

[24] Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for NLP. In: Proceedings of IEEE International Conference on Machine Learning, pp. 2790–2799 (2019)

[25] Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: QLoRA: Efficient finetuning of quantized LLMs. In: Proceedings of Advances in Neural Information Processing Systems, pp. 10088–10115 (2023)

[26] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Proceedings of Advances in Neural Information Processing Systems (2023)

[27] Mistral AI Team: Mistral Small 3 (2025). https://mistral.ai/news/mistral-small-3

[28] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)

[29] Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)

[30] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[31] Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., *et al.*: Searching for MobileNetV3. In: Proceedings of IEEE International Conference on Computer Vision, pp. 1314–1324 (2019)

[32] Tan, M., Le, Q.: EfficientNetV2: Smaller models and faster training. In: Proceedings of IEEE International Conference on Machine Learning, pp. 10096–10106 (2021)

[33] Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., *et al.*: Swin Transformer V2: Scaling up capacity and resolution. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 12009–12019 (2022)