

Residential Seismic Vulnerability Identification using Large Vision and Language Models

Meredith Lochhead, Greg Deierlein, and Iro Armeni

Department of Civil and Environmental Engineering, Stanford University, California.
mlochhea@stanford.edu

This manuscript is a preprint and has not been peer reviewed.

ABSTRACT

One- and two-story wood-frame housing is the most common dwelling type in the United States. While it generally performs well in terms of earthquake life safety, certain seismic vulnerabilities, including crawlspace (cripple wall) foundations, living spaces over garages, and masonry chimneys, have led to high financial loss and displacement in past earthquakes. Building-level analyses have demonstrated significant performance differences due to foundation, garage, and chimney-related vulnerabilities, yet most regional seismic risk assessments use a single fragility function, selected based on era, to represent all wood-frame housing. Capturing the impact of specific vulnerabilities and targeted retrofit measures at the regional scale requires sufficiently detailed building inventory data. However, existing data rarely describes such features. This study addresses that gap by leveraging street-level imagery alongside Large Vision and Language Models (VLMs) to systematically collect relevant building features. Results show that VLMs can reliably screen images and identify broadly visible building features. Furthermore, this study contributes a reusable methodology for developing and evaluating prompts through repeatable strategies such as sequential questions to guide model behavior, explicit uncertainty accommodation, and assessment of uncertain or incorrect predictions, among others. While foundation types are inherently difficult to classify due to occlusions in the imagery and limitations of exterior building views, using probability-based outputs and leveraging neighborhood similarities enable nuanced assessments that could improve our understanding of exposure and uncertainty at the regional level.

INTRODUCTION

One- and two-story wood-frame construction is the most common dwelling type in the United States, making up over two-thirds of the housing units in the state of California [1, 2]. In the context of seismic risk, single-family wood-frame housing is not likely to collapse due to ground shaking, and it tends to perform well in terms of life safety goals [3]. However, past earthquakes have caused significant damage to such houses. For example, the 1994 Northridge Earthquake (magnitude 6.7) resulted in over 9,000 housing units with restricted post-earthquake occupancy, displacing between 80,000 and 125,000 people [4, 5]. More recently, the smaller 2014 South Napa Earthquake (magnitude 6.0) caused restricted occupancy in 1,865 housing units [5], demonstrating that even moderate magnitude earthquakes can have large impacts on wood-frame housing. Repairable and irreparable damage in residential construction causes a series of downstream consequences, including significant direct economic losses, the displacement of families, and an associated reduction in the workforce due to prolonged displacements [4, 6]. These negative impacts collectively demonstrate that the seismic performance of wood-frame housing is important for understanding earthquake resilience.

Regional assessments can be used to quantify the seismic risk in a community and measure the impact of seismic mitigation strategies. In most regional risk assessments, the seismic performance of all wood-frame housing is represented using a single vulnerability model per era [7]. However, performance in past earthquakes has consistently demonstrated that much of the damage and financial loss in wood-frame housing is driven by the presence of certain seismic vulnerabilities. These include (i) raised foundations, which can collapse or cause the house to slide off the foundation; (ii) living spaces over garages, which can lead to soft story collapse; (iii) masonry chimneys, which pose falling hazards; (iv) hillside houses, which face torsion loads under seismic shaking; and (v) mobile homes, which can be displaced and require re-installation after an earthquake [1, 8].

Since the Northridge Earthquake in 1994, there has been substantial research on the seismic performance of wood-frame housing at the building scale, including the development of vulnerability models for detailed housing typologies that differentiate performance based on era, number of stories, foundation type, materials, retrofit status, and other features [9]. Such models have the potential to improve regional seismic risk assessments and quantify the impact of retrofit measures. However, their incorporation into regional models has primarily been impeded by a lack of available building inventory data.

Commonly used building inventory datasets, such as the National Structure Inventory and the Hazus Inventory, lack information on specific residential seismic vulnerabilities [10, 11]. In addition, while local tax parcel data is often used as an inventory data source, it can be inconsistent across jurisdictions, and since the original intent of parcel data is not to measure seismic vulnerability, many relevant features are not collected. Previous regional risk assessment studies have bypassed this inventory data limitation by randomly assigning detailed typologies at each location [12]. While this does allow an increase in the typology resolution of the study by using more detailed structural models, it does not meaningfully increase the fidelity of the study due to the random sampling in the inventory.

In the absence of sufficient data, an alternative method is to collect missing building features using imagery of the built environment. This method has attracted considerable research interest because imagery data is widely accessible across many locations; however, extracting building features from imagery typically requires task-specific computer vision methods or manual annotation. In the context of natural hazard risk, Convolutional Neural Networks (CNNs) have commonly been used to collect relevant building features from street-level imagery, as discussed in Section 2. However, there are three main limitations of using CNNs in this context. First, using CNNs requires specific training efforts and corresponding labeled image datasets for each building feature of interest. In the context of residential seismic risk, there are many possible features that influence seismic vulnerability, meaning that this approach requires substantial effort to develop a set of labeled images. Second, CNNs may have limited generalizability. Models trained on imagery from specific geographic regions may not transfer well to other regions with different architectural styles, construction practices, or street-level image characteristics, leading to reduced accuracy when applied in new and varied contexts and/or a need for extensive labeled image sets to counteract it [13, 14]. Finally, CNNs cannot be flexibly queried once trained, as the task they perform and the categories they output are fixed by the original training labels. This limits the ability of the user to interpret or steer model outputs.

Large Vision and Language Models (VLMs) represent a class of foundation models that have the potential to address each of these limitations. First, these models are trained on a large corpus of image and text data, creating connections across visual and linguistic concepts that have been shown to generalize to different tasks [15], hence not requiring any task-specific training data. Second, since they are not trained or fine-tuned using narrow, task-specific images, VLMs may be more generalizable across locations than CNNs, particularly in the zero-shot context, where models are applied directly without any task-specific fine-tuning or labeled examples [13, 14]. Finally, these models provide answers via natural language prompt querying, increasing users' ability to interpret and steer performance. Thus, VLMs have the potential to provide accurate, detailed information at the building level in generalized settings; however, to the authors' knowledge, no study has applied them in the context of identifying seismic vulnerabilities in single-family wood-frame housing.

This study explores the use of zero-shot VLMs to collect sufficient building inventory data to inform detailed regional seismic risk assessments of wood-frame housing. More specifically, this study leverages street-level imagery through Google's Street View Static API and the GPT-5 model through OpenAI's Responses API to identify residential seismic vulnerabilities and other building features that are required for downstream seismic vulnerability model selection.

This study makes three contributions. The first is a zero-shot prompting method that uses street-level imagery to extract the building features and foundation details that govern the seismic performance of single-family wood-frame housing. The method performs the same data collection task as a human reviewer with comparable accuracy and requires no task-specific training data or field inspection. Prompting strategies reported in the literature are consolidated and analyzed across multiple prediction tasks, yielding guidance on which prompt characteristics drive performance for a given building feature. The second is an approach for quantifying uncertainty in VLM predictions, which is shown to produce reliable confidence estimates and allows inventory uncertainty to be propagated through regional seismic risk assessments. The third is a prediction strategy that incorporates imagery of neighboring houses, which is shown to improve predictions for individual houses and increase prediction consistency across a neighborhood. Together, these contributions provide a reproducible framework for collecting detailed residential building inventory data at the scale required for regional seismic risk assessment.

RELATED WORK

Recent advances have expanded the use of computer vision in extracting building features from street-level imagery. In the context of seismic risk, computer vision is often used to inform downstream vulnerability model selection, primarily through overall building typology classification or feature identification. This section reviews CNN-based classification, the shift towards VLMs for inventory development, how model outputs represent uncertainty, and the use of neighborhood context to inform predictions.

CNNs for Seismic Risk Identification using Street-Level Imagery

Previous studies have used street-level imagery and CNNs for overall building typology classification, which provides a holistic building-level assignment to inform seismic behavior. Gonzalez et al. [16] represent one of the first systematic efforts to classify building typology for seismic risk using deep learning, based on a case study in Colombia. Other studies use and compare CNNs for building typology classification, including studies focused on construction practices in Norway [17], Chile [18], and Indonesia [19]. These studies show that CNNs can achieve accuracies above 80% but that accuracy can be contingent on building material (e.g., Ghione et al. [17] find that timber construction has higher accuracy than reinforced concrete).

Similar frameworks using CNNs and street-level imagery have been extended beyond overall building typology classification to identify individual building features associated with seismic risk, such as multi-family soft story construction [20, 21] and unreinforced masonry [22]. While many of these studies have relied on whole-image classification, this approach often struggles with accuracy given real-world noise and occlusions and may have difficulty locating a building in an image [23, 24]. To address this, more recent studies have employed instance segmentation to localize specific building components, such as parking garages or open ground floors, more reliably [24].

Beyond direct seismic risk applications, CNNs have been used to extract individual characteristics such as building use [25], facade material [26], building age [27, 28], and building height [29, 30]. To support both building typology classification and feature identification, Silva et al. [31] develop a labeled image database of street-level imagery through detailed field surveys of Alvalade, Portugal to support machine learning classification, including information about construction material, age of construction, and number of stories. While this dataset has supported other computer vision studies [32], it is not used in the current study because the specific building features of interest (raised foundation in single-family homes, etc.) are not present in Alvalade. However, the methods developed in the current study could be adapted to identify other relevant building features in the Alvalade dataset or similar datasets.

Comprehensive evaluation of seismic risk depends on the interaction of multiple building features, so CNNs have also been used in multi-attribute frameworks to identify several building features. Recently, Shourabi and Bakhshi [33] train and assess various CNNs to identify a suite of urban building features associated with seismic risk, such as occupancy, building type, plan irregularities, vertical irregularities, possible falling hazards, and severe damage, with the goal of automating the Federal Emergency Management Agency’s (FEMA) Rapid Visual Screening system. In addition, Kumar et al. [32] compare feature extraction models to identify construction material, construction age, number of stories, number of basements, and position within the block, among other features. Kumar et al. [32] highlight the potential of vision-based methods to collect features relevant for the Global Earthquake Model taxonomy and demonstrate that prediction quality for several features is improved by leveraging multi-view imagery of the target building. Additionally, the Building Recognition Using Artificial Intelligence at Large Scales (BRAILS) framework contains a set of pre-trained CNNs that are fine-tuned using supervised learning to classify features for building inventory development. The tool has classification modules for roof shape, occupancy, number of floors, and soft stories, as well as object detection modules for components such as garages, which can be used to inform vulnerability of the building stock to natural hazards [34, 35]. The BRAILS models focus on slightly different seismic vulnerabilities than those considered in this study (see Section 3); for example, the soft story detection focuses on larger multi-family housing, rather than raised foundations in single-family homes, and the garage detection module identifies garages generally rather than living space over garages specifically. These differences in detected features limit the usefulness of BRAILS as a direct comparison point for the method developed in this study.

VLMs for Building Feature Extraction using Street-Level Imagery

In recent years, VLMs have changed and expanded the use of computer vision techniques for assessing the built environment. Peng et al. [36] conduct a literature review of VLM-enabled street-level analytics, including the different research domains, image data sources, models, applications, and associated limitations. Several studies have explicitly targeted building feature collection for inventory development using VLMs and street-level imagery, employing a variety of open-source models. For example, Pan et al. [14] introduce a zero-shot workflow using models such as CLIP [15] and SAM [37] to generate both image-level and segment-level captioning as an alternative to the BRAILS framework. In addition, Yao et al. [38] use GroundingDINO [39] and BLIP [40] to classify wall materials by posing a series of questions.

In addition to open-source models, proprietary models have also been applied to street-level-related computer vision tasks, most commonly OpenAI’s GPT models. Zero-shot prompts have been used alongside street-level imagery to assess categorical building era and building energy efficiency through facade-level analysis [13, 41]. These previous studies develop prompts iteratively and evaluate model performance by requiring the model to explain its reasoning.

In a related effort, Raghu et al. [42] introduce URBAN-AI, a workflow that applies several models (e.g., GPT-4V) to assess facade materials and other downstream indicators, including seismic risk measures in San Francisco, CA. However, this work detects different seismic vulnerability metrics than those considered in the current study, such as the presence of masonry, large openings, and multi-story buildings, which are more relevant for typologies beyond single-family homes.

Studies have also employed fine-tuned VLMs, which use additional labeled data to target specialized tasks such as building condition assessment [43], post-earthquake structural damage assessment [44], building feature collection such as height, usage, and material [45], and roof material identification, which is important for risk across multiple natural hazards [46].

Overall, these studies demonstrate that VLMs can be used across a variety of applications to assess urban data, collect building features, and inform energy and natural hazard risk. This existing VLM literature is used to inform the initial prompt format, the use of structured outputs, prompt iteration, and model reasoning in this study. However, these studies generally have not designed multi-prompt workflows, nor do they incorporate systematic uncertainty quantification or visual neighborhood-level context. Finally, while there has been substantial work on detecting building materials and general attributes, the authors are not aware of any study that has focused specifically on a multi-feature assessment of single-family residential seismic vulnerability using VLMs, which is the goal of this study.

Probability-Based Outputs and Uncertainty in Vision-Based Inventory Development

Vision-based models typically produce point predictions, with a single label or category assigned. This mirrors the use of a single building inventory in a regional seismic risk assessment, despite uncertainty in the underlying data.

One potential source of uncertainty comes from the human-generated output labels used to train or evaluate a model. Pelizari et al. [18] note that because labels stem from human interpretation, machine learning systems carry epistemic uncertainty that should be propagated through seismic risk modeling. Several studies account for this uncertainty in annotations by asking human labelers to provide soft labels, probabilities, or confidence scores that can be used in model training to improve robustness [47, 48].

Other studies discuss multi-label or probability-based approaches to accommodate uncertainty and task complexity. For example, Raghu et al. [26] use multi-label classification to account for real-world complexity, such as building facades with multiple materials; however, multiple outputs in this case do not indicate uncertainty, but rather the actual presence of multiple materials. Zeppelzauer et al. [28] compare majority voting and an overall likelihood strategy for building age assignment, where class likelihoods of individual image patches are averaged to provide a more sensitive global estimate of building age. More recent research involving multi-view imagery collects probability scores from different viewpoints of the same building and combines them to determine a final, most-probable class [32]. However, these studies typically use probabilities as intermediate outputs, rather than explicitly as a tool to denote uncertainty in final model predictions. Yan and Huang [49] measure uncertainty by providing a variance term in addition to the estimated height of a building. Law and Miura [46] do not use probability-based outputs, but allow for model abstention, where the model abstains from making a prediction in cases with ambiguous visual cues.

Outside of building feature assignment, probabilistic outputs and model uncertainty in VLMs have been explored more broadly, primarily in cases where (i) confident, incorrect answers are costly, or (ii) there is genuine ambiguity in the output. Previous studies on VLM uncertainty have used conformal prediction methods; however, these rely on token-level probabilities, which are often not available in proprietary models [50, 51]. Previous literature has also explored methods to output model confidence, such as asking the same question multiple times and assessing variation across outputs, or prompting for linguistic likelihood expressions or numerical probabilities as part of the model output. The latter strategy is referred to as ‘verbalized’ confidence or uncertainty, and is found to be better-calibrated with respect to model correctness [52]. Verbalized probabilities have been explored as output weights summing to one, including previous efforts to fine-tune large language models using reinforcement learning to calibrate verbalized probabilities [53]. The current study builds on existing work by exploring the use of verbalized probabilities in the zero-shot VLM context to account for uncertainty in model predictions and allow inventory uncertainty to be propagated through to seismic risk estimates.

Use of Neighborhood Context in Vision-Based Inventory Development

Previous vision studies for building feature extraction typically use one or more images of the house of interest. The current study explores the impact of additionally using images of neighboring houses to inform building feature assignment. While the authors are not aware of explicit examples of this in the literature, several sources highlight the potential of using spatial or neighborhood context to refine predictions [35, 54]. As part of the BRAILS framework, spatial patterns in building metadata are explicitly leveraged to map information from nearby buildings to a target

building with missing data (i.e., data imputation); however, these methods typically rely on statistics from tabular data, rather than imagery of neighboring houses. Ghione et al. [17] propose providing CNNs with pre-conditioning data about the neighborhood to increase classification accuracy as a direction for future work; however, this approach is not explored further. Finally, the BuildingView framework incorporates the presence of neighboring buildings as a descriptive attribute, but does not use their visual characteristics in predicting the features of the target building [55]. Together, these studies suggest that neighborhood context is promising, but the use of neighboring building imagery for feature extraction has not been explored.

REQUIRED BUILDING FEATURES

This study aims to identify three main sources of seismic vulnerability in single-family homes: raised foundations, masonry chimneys, and living spaces over garages. These vulnerabilities are the focus because they have well-established, relatively affordable prestandard seismic retrofit plans from FEMA P-1100 and active retrofit programs through the California Earthquake Authority’s (CEA) Residential Mitigation Program [1, 8]. Mobile homes, which also have an active retrofit program, are excluded from this study as street-level imagery often does not cover mobile home parks. This study also considers secondary features that support foundation type identification, as well as additional building features required to select an appropriate vulnerability model. All features are summarized in Table 1, and further detail on the seismic vulnerabilities follows.

TABLE 1. Building Features Collected from Street-Level Imagery

Feature	Description
<i>Primary seismic vulnerabilities</i>	
Foundation Type	Classified as crawlspace, stem wall, or slab-on-grade. Raised foundations (i.e., crawlspace and stem wall foundations) have been shown to drive damage in past earthquakes. Slab-on-grade foundations are typically less seismically vulnerable [9].
Masonry Chimney	Masonry chimneys, which can be classified as either exterior or interior chimneys, have the potential to crumble or collapse during an earthquake. This can pose a falling hazard to the occupants or cause roof or other structural damage [1].
Living Space over Garage	An occupied upper story extending over a ground story garage can pose a seismic risk. The large open space can function like a soft story if the walls are not sufficiently braced [1].
<i>Secondary features supporting foundation type identification</i>	
Foundation Height	Crawlspace foundations are typically the tallest and can vary in height, followed by stem wall foundations; slab-on-grade foundations sit essentially at grade. Thus, the height of the first floor above grade can help distinguish foundation type and also affects the seismic performance of crawlspace foundations.
Entry Stairs	Stairs leading to the front door often indicate a raised foundation. Crawlspace foundations are typically characterized by three or more stairs.
Foundation Vents	Vents along the lower exterior portion of the wall indicate ventilation under a raised foundation. These may be associated with crawlspace or stem wall foundations, but not slab-on-grade.
Lot Slope	Lot slope, identified from the grade change directly beneath the building footprint, indicates whether the house sits on a flat or sloped lot. Houses on high-slope lots fall outside of the standard plan set and require engineered, site-specific retrofit designs [1].
<i>Additional building features for vulnerability model selection</i>	
Number of Stories	The number of stories strongly influences the expected seismic performance of a house [5].
Roof Material / Roof Weight	The roof material (e.g., shingles, ceramic tiles) influences seismic performance primarily through roof weight [9].
Siding Material	The siding material (e.g., stucco, wood) affects the lateral stiffness of the exterior walls. This is particularly important for the performance of the crawlspace walls [9].
Year of Construction	The year of construction reflects relevant practices and materials at the time the house was built, as well as eligibility under seismic retrofit program cutoffs [8].

Crawlspace foundations (i.e., cripple wall foundations) have short wood-framed walls connecting a concrete foundation to a wood-framed first floor, forming a crawlspace [1, 9]. In older houses, the exterior cladding may provide the only lateral resistance in crawlspace walls. Such unbraced walls have been major contributors to earthquake damage [56, 9]. Stem wall foundations have first floor framing resting directly on raised concrete or masonry perimeter walls [1, 9]; in some cases, the siding stops above the top of the concrete or masonry wall, making it visible from the exterior. Both crawlspace and stem wall foundations may have inadequate sill anchorage, which allows houses to shift or slide off their foundations during shaking. Slab-on-grade foundations, which became increasingly popular after 1960, sit at ground level; they typically include anchor bolts when built to code, substantially reducing the seismic vulnerability [57]. Schematics of the two raised foundation types are shown in Figure 1.

The CEA Earthquake Brace and Bolt program specifically targets pre-1980 houses with raised foundations (crawlspace or stem wall), as houses built more recently typically have sufficient bracing to withstand seismic loads. Furthermore, this program only targets houses on flat lots, rather than high-slope lots, as hillside houses face additional torsional demands and thus require engineered retrofit designs [1].

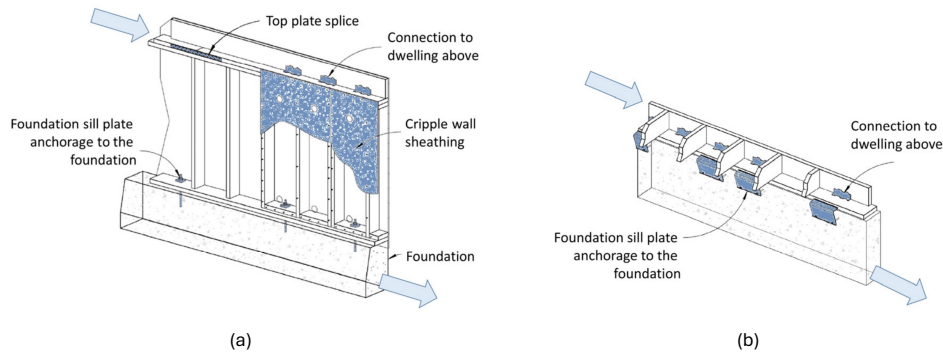


Fig. 1. Schematic of seismically vulnerable raised foundation types: (a) crawlspace and (b) stem wall foundations. Figure from Welch and Deierlein [9].

Masonry chimney failures were reported following the 2014 South Napa earthquake, including both toppling of entire chimneys and crumbling of masonry [56]. Masonry chimneys can be differentiated into exterior chimneys (one wall of the chimney can be seen on the dwelling exterior for at least one story height) and interior chimneys [1], which have different retrofit recommendations and failure modes. Non-masonry chimneys do not pose the same seismic risk.

Elevated seismic vulnerability may exist when an occupied living space in an upper story extends substantially or completely over a ground story garage [1]. Living-space-over-garage houses built prior to 2000 may have insufficient lateral strength and are targeted by the CEA Earthquake Soft Story retrofit program [8].

IMAGE SET DEVELOPMENT

This study uses street-level imagery available through Google’s Street View Static API for all data collection. The API returns an image based on a location and selected parameters such as *heading* and *field of view*. Images can be downloaded based on point locations of buildings; however, building footprint polygons are commonly used as the baseline geometry of building inventories because they enable robust linking of data across sources and physically represent single buildings [58]. To remain consistent with other inventory development efforts, this study uses a method based on building footprint polygons to obtain street-level imagery.

Image Download Pipeline

BRAILS++ is a commonly used framework for using street-level imagery to assess building features [34]. This framework obtains street-level imagery based on building footprint polygons by retrieving the closest available view to the building centroid. However, a single image does not always provide a clear view of a given building. For example, there may be occlusions in the image, the nearest view may be looking at the side of a house (particularly for buildings on street corners), and the line of sight may pass through other buildings, producing misleading views. Furthermore, not all seismic vulnerabilities are visible in every view; for example, a masonry chimney may be visible in an oblique view of a house, but not a head-on view. To address these limitations, this study develops an extended image download pipeline that retrieves multiple street-level images per footprint.

First, eight points are placed equidistantly from the footprint centroid in a circle with a radius of 15 meters, shown in blue in Figure 2. Second, for each of those eight requested points, the nearest available outdoor Street View panorama ID is found, as shown in green in Figure 2. There may not be a 1:1 mapping between the requested points and available panoramas, because multiple requested points may map to the same nearest available panorama. The heading for the image download is then calculated as the direction from each available panorama to the centroid of the footprint. The distance from the panorama location to the closest edge of the building footprint is also recorded, and based on that distance, the field of view is adjusted slightly (e.g., a wider field of view for a closer distance). Using the panorama ID, heading, and field of view, between zero and eight images are downloaded for each building footprint, depending on how many unique panorama IDs are available. The eight-image cap is a result of originally placing eight points around the footprint to search for available panoramas.

Based on initial observations, three filters are developed to prevent the use of poor-quality or misleading images in downstream vision predictions. These filters are developed by examining downloaded images, but are ultimately applied before the API call to improve the quality of images being passed into downstream vision models and prevent the unnecessary download of additional images. These filters are shown in Figure 2, and all pertain to the line of sight between the panorama location (camera location) and the nearest point on the building footprint. First, the *distance filter* removes images where the length of the line of sight exceeds 33 meters, shown in Figure 2(a). Empirically, most images that exceed this distance limit do not offer clear views or show multiple buildings. Second, the *intersection filter*, shown in Figure 2(b), removes any images that have their line of sight intersecting another building footprint. Some of these images have clear views of buildings; however, due to the intersection of the line of sight with another footprint, these images show the wrong building. Finally, the *line of sight filter* removes cases that appear to have a clear line of sight (i.e., not flagged by the intersection filter), but in reality do not. For example, the house shown in Figure 2(c) has a clear line of sight past the surrounding footprints; however, the resulting image does not primarily show the building of interest, instead focusing on a neighboring building in the foreground. To eliminate these types of images, cases where the line of sight comes within 0.66 meters of another building footprint are also removed.

There are two possible reasons why building footprints may not have any available imagery. First, the Street View Static API may not return any panorama IDs close to the point of interest. Observations show that this typically occurs in newly developed areas that may not have street-level availability, as well as on privately owned land, such as apartment complexes or mobile home parks. This study focuses on seismic risk in older single-family wood-frame homes, so these gaps in the available imagery are not consequential in this case. Second, footprints also may not have imagery due to the three aforementioned filters, such as the footprint shown in Figure 2(c).

This image download pipeline does not account for cases where the street-level imagery was captured before a demolition, reconstruction, or renovation, so the VLM-informed inventory may reflect outdated building conditions. Because older buildings carry higher seismic risk, this could lead to overestimated seismic risk, especially in neighborhoods with high reconstruction rates and infrequently updated imagery.

Labeled Image Sets

This study explores the use of zero-shot VLMs for predicting building features, and thus does not require labeled training images. However, labeled images are used to develop appropriate prompts and evaluate model performance and generalizability across locations. Thus, the proposed image download method was used to curate several datasets for prompt development and model evaluation. Though methods developed in this study are intended to be generalizable across California, a primary case study city of Hayward, CA was selected. Hayward consists of predominantly single-family residential construction and the Hayward Fault runs through the city, making it an appropriate case study for quantifying residential seismic risk.

A set of 345 building footprints was selected at random across the city of Hayward, and the associated $\sim 1,400$ images (1–8 images per footprint, per the above method) make up the *prompt development image set*, which is used to evaluate and select appropriate prompts. Any footprint with no available imagery was excluded from the prompt development set. Ground truth labels are required to evaluate model accuracy and compare prompting strategies. Thus, all relevant features (see Table 1) other than foundation type were hand-labeled for the prompt development set by an expert. Because most houses have multiple complementary images, building feature labels are assigned at the *building* level, not at the *image* level. The year of construction was obtained from the Hayward tax parcel data, as described in Lochhead et al. [58].

Many building features considered in this study are generally easy to identify from the exterior of the house (e.g., number of stories and siding material). However, this is not always the case. In particular, differentiating between foundation types can be challenging for two reasons. First, occlusions in the imagery such as cars, fences, and



Fig. 2. Image download filters with example images. (a) Distance filter, where the building of interest is too far away; (b) Intersection filter, where the image is clear but does not show the building of interest; (c) Line of sight filter, where the line of sight passes too close to another footprint, creating a misleading view.

landscaping are particularly obstructive for seeing the foundation of the house. Second, even if there are no occlusions in the image, it can sometimes be difficult to differentiate between foundation types using only the exterior views of the building. Thus, the ground truth data for foundation type classification was developed differently from the other ground truth building feature labels. To accommodate both the occlusions and the general lack of certainty about foundation types using only the exterior views, the ground truth foundation type assignment allows for multiple labels to be assigned as possible foundation types of each house, indicating uncertainty. For example, if it is unclear whether a house has a crawlspace foundation or a stem wall foundation, both tags can be associated with the building footprint.

Furthermore, to reduce subjectivity in ground truth assignments, three experts independently labeled possible foundation types for the prompt development set. Each labeler could select one or more possible foundation types from a pre-defined list. If the foundation type was fully unclear from the images, labelers selected all possible foundation types for that building. If at least two of the three experts selected the same set of labels for a given footprint, that set was adopted as the ground truth, regardless of how many foundation types it contained. If no two labelers agreed on the same set, the ground truth was assigned as the union of all foundation types given for that building, on the assumption that this level of disagreement indicates genuine ambiguity.

In addition to the prompt development image set, four *test image sets* were developed to evaluate final model performance and generalizability. The images in the prompt development set are all from the city of Hayward; however, the model developed in this study aims to be relevant across the state of California. Thus, the test sets each consist of 130 footprints selected from four different cities. The first test site is Hayward, which is used to evaluate final model performance on the same city its prompt was developed for. The Hayward test set was developed to have no overlapping buildings with the prompt development set. San Bernardino has relatively similar makeup to Hayward in terms of foundation types and seismic risk, and it serves as a second test site. Older neighborhoods in Oakland and Los Angeles were selected as the third and fourth test sites due to their high concentration of older houses with crawlspace foundations, which differentiates them from other sites and results in high levels of vulnerability. The building footprints associated with each test site were sampled randomly, as in the prompt development set; thus, the test sets were not developed to have an even split between categories, but rather to be reflective of their respective locations. This was done to better capture actual model accuracy when deployed in a given city, rather than on a

controlled test set. Figure 3 provides an overview of the prompt development and test image sets used in this study, and Figure 4 provides several representative images for each location.

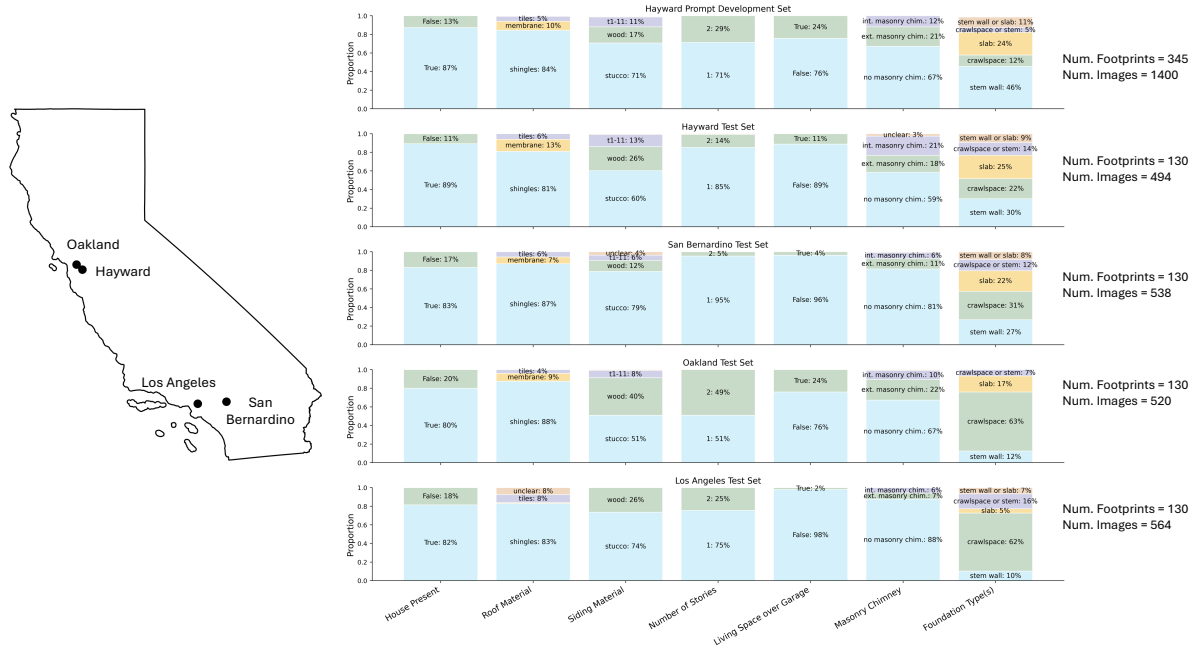


Fig. 3. Overview of prompt development and test image sets. Proportions of each label by feature in the ground truth data. For features other than ‘house present,’ proportions are shown for buildings labeled as single-family homes.

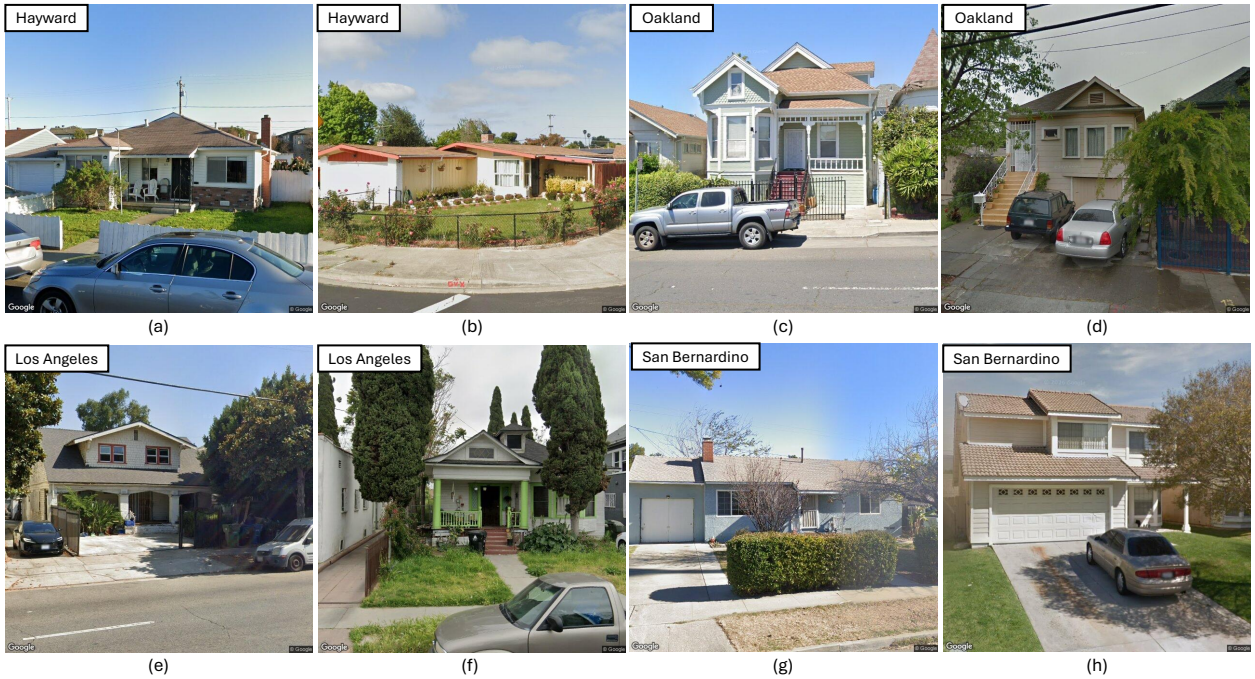


Fig. 4. Example images across four test locations.

MODEL SETUP

All prompting is done using OpenAI’s Responses API with structured outputs, which is a commonly used method to enforce a JSON schema for model outputs. Model selection and inference details are described in Section 11. Before developing formal prompts, the API was used to probe the model informally to assess baseline understanding of relevant building characteristics. This involved asking questions about materials and other features, their appearances, and how these features might be identified in street-level imagery. These informal tests confirmed that the model had sufficient conceptual and visual understanding of the features of interest.

Initial model testing indicated that collecting all required features of interest within a single prompt degraded model accuracy due to the prompt length and complexity. Thus, to improve overall performance, the prompt was split into three distinct tasks: 1) screening, 2) building features, and 3) building foundation details.

Per the image download pipeline discussed in Section 4, between one and eight street-level images are available for each building. First, all images are passed into the screening prompt, which assesses whether the building is a single-family home (and thus should be evaluated further), then identifies and removes views that do not contain a usable image of the building of interest (e.g., the house is obstructed by vegetation). Only images that pass screening are passed into the two subsequent prompts.

Several experiments were conducted for the building feature and foundation detail prompts comparing the accuracy of predictions given: 1) the single closest image, 2) multiple images per building, each evaluated separately by the prompt, and 3) multiple images per building passed into the prompt together and evaluated as a set. Empirical testing showed a clear benefit to passing multiple images together in a single prompt, consistent with findings in previous literature [32]. Thus, all evaluation and accuracy calculations for the building feature and foundation detail prompts are done at the *building* level, not at the *image* level.

A schematic of the full three-part prompt setup is shown in Figure 5, including an example of the screening prompt applied separately to each image before the feature and foundation prompts are applied to the filtered image set together. Figure 5 also details which specific building information is collected by each of the three prompts.

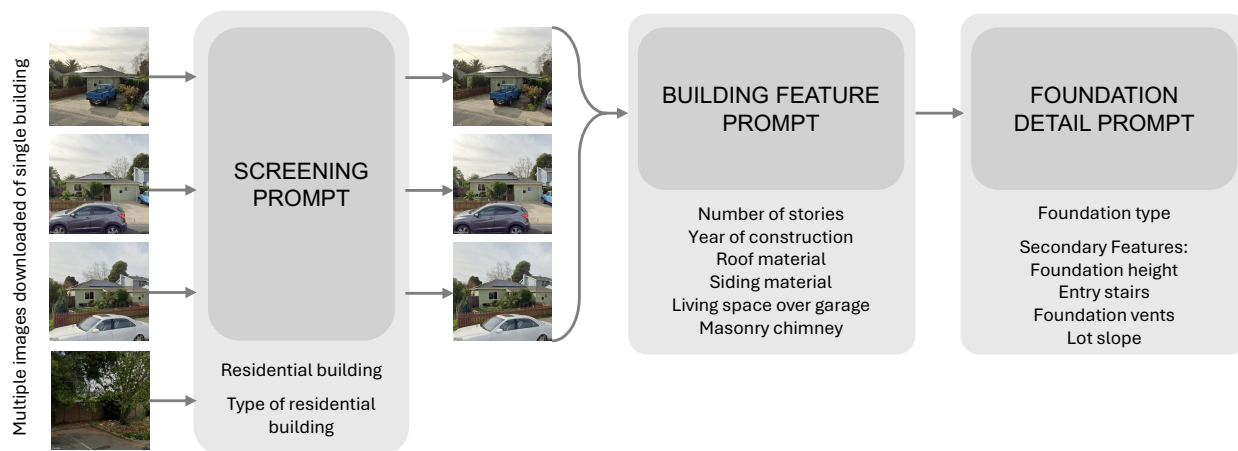


Fig. 5. Overview of three-part prompt setup. Multiple images per building are screened individually, then subsequently passed together into the two downstream prompts.

The three prompts are all structured in a similar way. First, each prompt identifies the role of the model and provides an overview of the task. For example, ‘you are a building contractor who is evaluating street-level images of a house to classify its features’ or ‘you are a structural engineer who is evaluating street-level images of a house to classify the foundation type for a seismic risk assessment.’ After this role and task definition, more explicit instructions are given, such as ‘use the guidelines below to identify the features of the building shown in the images provided. Classify features that are clearly visible in the images. Do not speculate.’ After these instructions, each prompt has a ‘Features to Classify’ section which includes a numbered list of required output feature formatted with the field name, data type, and possible output options with corresponding descriptions. These same fields are specified as required outputs in the structured output schema passed to the Responses API, ensuring consistent fields across all predictions.

Starting from this overall prompt setup, initial prompt iteration was conducted by changing the language describing

the task, the descriptions of the possible output features (e.g., visual description of stucco), and the overall level of detail used in the prompts. This iteration resulted in functioning prompts with clear, concise descriptions of tasks and materials; this is referred to in the remainder of the study as the ‘initial prompt.’ Beyond updating wording and descriptions, more detailed prompt development and evaluation were conducted independently for the screening, building feature, and foundation detail prompts using the prompt development image set. More specifically, this study identifies several prompt development strategies and evaluates their impact on performance across prediction tasks.

- **Iterative prompt refinement using model-reported reasoning.** VLMs can be prompted to provide explicit reasoning for a classification. Reasoning can be reviewed to understand the basis for the prediction and identify misinterpretations or ambiguity, and prompts can then be revised accordingly. This approach is generally consistent with iterative prompt engineering strategies and human-in-the-loop refinement [59].
- **Sequential questions to guide model behavior.** As opposed to posing a single question, prompts can instead pose a sequence of targeted sub-questions to guide the model through relevant features before producing a final prediction. This is similar to chain-of-thought prompt approaches, in which decomposing a task into ordered steps improves performance on complex tasks [60].
- **Explicit uncertainty accommodation.** An ‘unclear’ category can be included in the output label set, allowing the model to abstain from a definitive classification in ambiguous cases. Forcing a prediction when image quality or visibility is poor can introduce systematic error, and there are recognized overconfidence issues with VLMs [61]. Accommodating uncertainty with a categorical label or allowing for model abstention has been used in previous classification tasks to reduce overconfident, incorrect predictions [62, 63, 46]. In this case, it is preferred to have the model select ‘unclear’ rather than make an incorrect or speculative prediction because uncertainty in the building data can be measured and propagated through a regional seismic risk assessment, whereas incorrect predictions may cause bias in the results.
- **Iterative prompt and label refinement using patterns in uncertain and incorrect predictions.** Unlike the reasoning-based review above, this strategy uses error patterns across the image set to identify systematic issues (e.g., visual or label ambiguity). Uncertain and incorrect predictions can be reviewed to identify recurring failures, and prompt wording and output labels can be updated accordingly.
- **Engineering context to address ambiguity.** A benefit of VLMs is the flexibility to provide context in prompts beyond a simple task description. Ambiguity in the task or image can be a source of error, and prompts can be revised to include additional contextual information, such as relevant definitions, domain knowledge, or task framing, to constrain the model’s interpretation.
- **Numerical versus categorical outputs.** For tasks involving numerical outputs, such as year of construction, prompts can be formulated to provide either direct numerical estimates (e.g., year of construction is 1973) or categorical estimates (e.g., 1970-1990 era). A prior application of a VLM for assessing year of construction uses categorical eras [13].
- **Few-shot prompting.** Example images and their corresponding labels can be included in the prompt to provide task-specific examples. Few-shot prompting can reduce ambiguity and has been shown to improve VLM output quality [64, 65].

Three additional strategies are identified and explored specifically in the context of the foundation detail prompt.

- **Use of probability-based model outputs with various postprocessing methods.** Rather than predicting a single output label, the model can instead be required to return weights (summing to one) associated with each possible foundation type. Results can then be postprocessed in various ways, such as assigning the maximum probability foundation type or selecting multiple possible foundation types based on outputs. This strategy allows for identification and propagation of uncertainty through a downstream regional seismic risk simulation.
- **Use of neighboring images to inform predictions.** In suburban areas in the United States, neighborhoods are commonly developed as part of subdivisions, where a large tract of land is zoned and developed by a single developer or a set of developers. Particularly in the post-WWII era of suburban growth, many homes were built simultaneously and with similar designs due to the lower per-unit costs of standardized designs and bulk material purchases, leading to design and construction similarities in neighboring houses [66, 67]. In light of this, it is possible that images of neighboring houses could provide helpful context in assessing the foundation type for a given house, and this is explored as a strategy for improving performance.
- **Prompting multiple times using the same prompt, then taking majority vote.** A previously used approach

to improve robustness is self-consistency decoding, where the model is sampled multiple times for the same input and the final prediction is selected based on the most frequent output across samples [68].

EVALUATION METRICS

Several evaluation metrics are used throughout this study for prompt development and assessment. For screening and building feature identification, the primary metric is overall percent accuracy. Many building features evaluated in this study (e.g., siding material) inform downstream model selection but do not represent seismic vulnerabilities themselves. Accuracy is therefore an appropriate metric, as no single class carries disproportionate consequences for downstream analysis.

For features directly indicating seismic vulnerability, such as the presence of a masonry chimney, both accuracy and false negative rate (the rate at which ground-truth positive cases are incorrectly classified as negative) are considered. False negatives directly translate to missed seismic vulnerabilities and thus should be minimized in prompt development.

Percent accuracy is not a straightforward metric for the foundation prompt due to the presence of multi-label ground truth values (e.g., a true label of "stem wall or crawlspace"). Model performance is therefore assessed across four complementary accuracy metrics, as shown in Table 2. Together, these metrics reveal the tradeoff between broader coverage through uncertainty and more specific predictions at the cost of overall accuracy. Because the three foundation types carry different levels of seismic vulnerability, per-class recall is also computed for each type, both with and without counting uncertain predictions. This allows for foundation-type-specific insights and the selection of a model that is most sensitive to the highest-risk foundation types.

TABLE 2. Description of four complementary accuracy metrics used to evaluate foundation type predictions. *Overall accuracy* refers to the combined total across the four metrics.

Description	Example	
	True Label	Predicted Label
Percent of predictions exactly matching the true label	"stem wall or crawlspace"	"stem wall or crawlspace"
Percent of predictions encompassing the true label (2 foundation types)	"stem wall"	"stem wall or crawlspace"
Percent of predictions encompassing the true label (due to uncertainty: prediction of all foundation types, or "unclear")	"crawlspace"	"unclear"
Percent of predictions capturing a subset of multiple true labels	"stem wall or crawlspace"	"stem wall"

Finally, all building feature and foundation prompts are evaluated independently of screening, meaning errors from the screening prompt are not propagated as inputs into subsequent prompts. This allows for the performance of each prompt to be assessed in isolation, independently of upstream errors.

PROMPT DEVELOPMENT FOR SCREENING

The screening prompt serves two purposes. First, it screens based on occupancy class; only single-family homes require further evaluation, so images of large apartment buildings, commercial buildings, or other non-single-family occupancy classes are removed at this step. Second, it filters out images with severe occlusions (e.g., trees, cars, and fences) that are not sufficiently clear to support further assessment.

All screening prompts evaluated in this study follow the overall prompt setup described in Section 5. The initial prompt addressed both screening tasks simultaneously by asking only whether a house was present in the image. While prediction accuracy with this approach was relatively high, a number of false positives reduced overall accuracy. *Patterns in uncertain and incorrect predictions* revealed that most false positives corresponded to images of mobile homes and multi-family homes, or images with intentional blur in which a portion of a house remained visible.

To address these errors, we evaluated the impact of using *sequential questions to guide model behavior* in the screening prompt. The updated prompt splits the screening task into three sequential steps: first, the prompt asks if there is a residential building in the image, then asks the model to specify the occupancy class (mobile home, multi-family, single-family), then finally asks if there appears to be an intentional blur on the image. Model outputs are handled in two different ways. First, the presence of a residential building and intentional blur are image-level characteristics, so individual images are filtered out. Conversely, occupancy class should be consistent across all images of a given building, and is therefore assessed at the building level to assign a binary label indicating whether the building qualifies as a single-family home. Assignment is based on the proportion of images falling into each occupancy class, and threshold values (0.67 for multi-family and 0.5 for mobile homes) were intentionally set to bias

ambiguous cases toward single-family homes, as they are the focus of the study. These thresholds were determined empirically using the prompt development image set and were found to generalize well across test cases.

Figure 6 shows that the more detailed screening prompt outperforms the simpler approach, demonstrating the benefits of *using patterns in uncertain and/or incorrect predictions* and of *sequential questions to guide model behavior*. Both prompts exhibited similarly low false negative rates, indicating that neither frequently missed residential buildings, and therefore neither is likely to overlook the associated seismic risk. However, the detailed screening prompt nearly doubled the number of cases correctly predicted as not being single-family homes. This matters because misclassifying such buildings would cause downstream risk assessment to assign inappropriate vulnerability models, and the buildings could be inadvertently included in the modeled impacts of a potential retrofit program, biasing results. The more detailed screening prompt was therefore selected on the basis of higher overall accuracy.

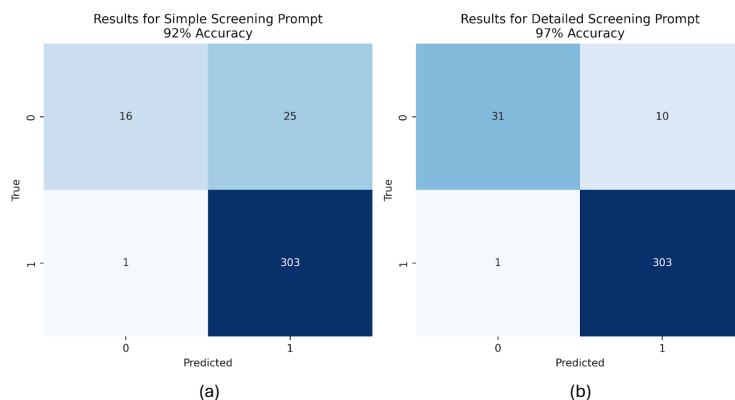


Fig. 6. Confusion matrices of screening predictions for (a) the simple prompt and (b) the more detailed prompt.

PROMPT DEVELOPMENT FOR BUILDING FEATURES

The building feature prompt is used to collect information for all buildings that have at least one image that passes the screening step. In this prompt, all screened views of a given building are passed into the prompt together for evaluation, as shown in Figure 5. The building feature prompt collects information on the number of stories, roof material, siding material, living space over garage, masonry chimney, and year of construction.

All building feature prompts evaluated in this study use the same setup (role and task definition, followed by a list of features to classify with descriptions), as described in Section 5. Several building features were already classified with high accuracy, given the initial prompt, including the number of stories (99% accuracy) and the presence of a living space over garage (95% accuracy). These features were both visually obvious and unambiguous, and thus may have been easier for the VLM to extract from the imagery. Other features had more variable performance in the initial prompt and required additional prompt iteration.

Roof Material

In the context of seismic risk, roof material is primarily used to inform the roof weight. In California, the majority of roofs are either shingles, which map to the ‘light roof’ designation, or curved ceramic tiles, which map to the ‘heavy roof’ designation [9]. Figures 4(a) through 4(g) show examples of shingles with various colors and textures, and Figure 4(h) shows an example of ceramic tiles.

In some images in the prompt development set, the roof material is not always visible due to the angle of the camera or other occlusions. In the ground truth labels, these cases are labeled as ‘unclear.’ Thus, the roof material prompts include an ‘unclear’ label (*explicit uncertainty accommodation*).

Assessing *patterns in uncertain and incorrect predictions* revealed two systematic errors. First, the model was over-predicting ceramic tiles; nearly all incorrect tile predictions were actually large-format shingles (see Figure 7(b)), which visually resemble ceramic tiles, but map to the ‘light’ roof weight designation instead of ‘heavy’. In addition, the model was over-predicting the number of roofs assigned ‘unclear’ relative to ground truth. These incorrect predictions were primarily low-slope or flat roofs (see Figure 7(b)). In response, the possible model outputs were modified to have two additional categories: large-format shingles and low-slope membrane roofs, both mapped to the ‘light roof’ weight

designation. The prompt was further adjusted using *engineering context to address ambiguity* by specifying that most low-slope or flat roofs are typically membrane roofs, even when the material itself is not visible.

Modifying the prompt and output labels increased roof material accuracy by 14 percentage points and roof weight accuracy by 16 percentage points, demonstrating the effectiveness of the prompt development strategies. Figure 7(a) shows the confusion matrices using the initial prompt, Figure 7(b) shows examples of the two roof types and how they were misclassified, and Figure 7(c) shows the updated output labels and roof weight mapping.

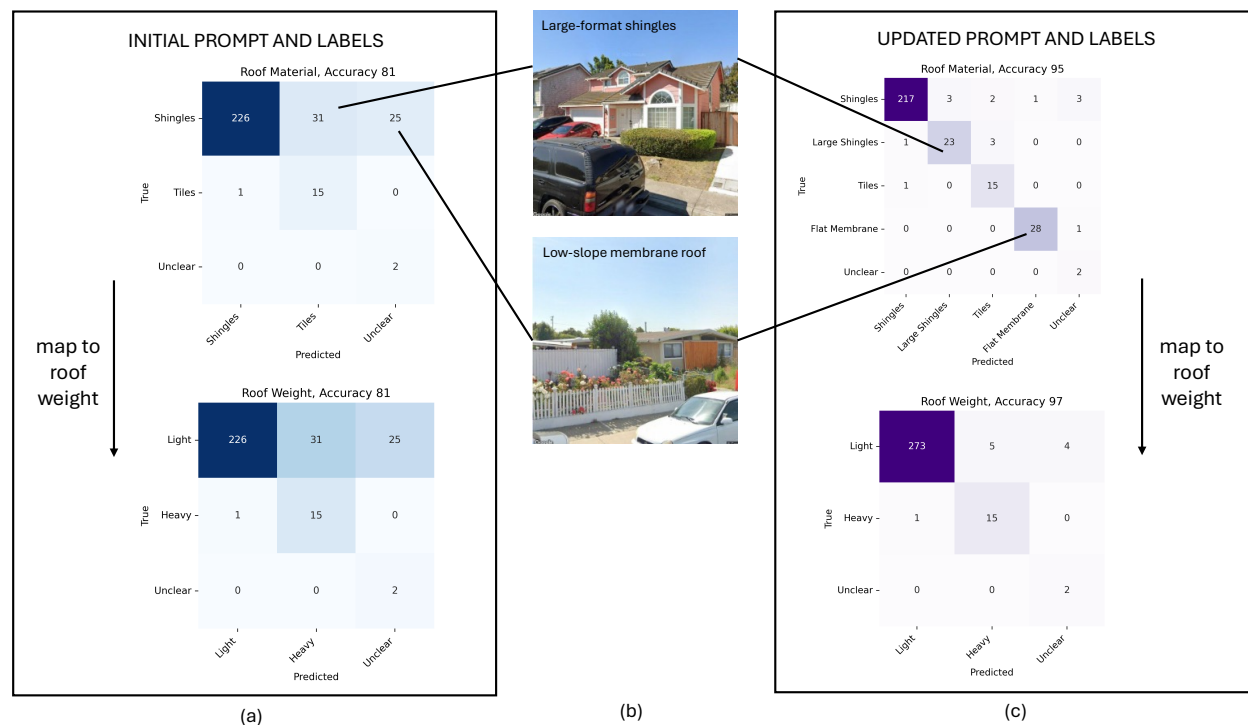


Fig. 7. Overview of roof material prompt development. (a) Results of initial prompt, with incorrect labels identified as large-format shingles and low-slope membrane roofs. (b) Examples of misclassified roof types. (c) Results of updated prompt, which accounts for these categories and improves roof material and weight accuracy.

Siding Material

In suburban California, common siding materials include stucco, wood, and T1-11 (wood panels with vertical, evenly spaced grooves; Figure 4(b)). Brick, stone, and other materials also appear, but typically as architectural details rather than primary siding (Figure 4(a)). Some houses also have mixed siding materials, with multiple materials each covering a substantial portion of the facade.

Given this, two labeling approaches were considered for the ground truth data: labeling only the primary siding material (using engineering judgment when multiple materials are present, based on which is most reflective of lateral stiffness), or recording all significantly present siding materials. While the latter is more comprehensive, existing seismic vulnerability models only support single-material inputs [9], so ground truth labeling focuses on the primary siding material to match its downstream use. As with roof material, *explicit uncertainty accommodation* is used to handle cases where the siding material is unclear, allowing an ‘unclear’ label in both ground truth and predictions. Unclear cases are less common for siding than roof material, since siding visibility depends less on camera angle.

The initial prompt performed well for houses with fully stucco or wood siding. However, *patterns in uncertain and incorrect predictions* showed that the initial prompt led to 1) misidentification of T1-11, 2) over-identification of materials used for architectural details as siding materials, and 3) low accuracy for mixed-material houses.

Regarding low T1-11 accuracy, we hypothesized it might be underrepresented in VLM training data relative to wood and stucco, and thus may benefit from *few-shot prompting*. Including a labeled example of T1-11 siding marginally improved performance (3 percentage point increase in accuracy). However, we concluded that this marginal increase in accuracy did not justify the additional cost and complexity in the prompt, and thus, this strategy was not used further.

To address over-prediction of architectural-detail materials (e.g., brick or stone chimneys and accent walls) and lower accuracy on mixed-material houses, we tested *sequential questions to guide model behavior* and *engineering context to address ambiguity*. First, the model was prompted to identify all siding materials, and then was specifically asked about which materials were most important for the lateral stiffness of the house. By structuring the output as a two-step reasoning process, we hypothesized that this would reduce conflation of architectural details with primary siding material, while also providing a more technical basis for selecting a material from a mixed-material house. However, the sequential questions increased prompt length and complexity, while resulting in only a 1 percentage point gain in overall accuracy. Thus, this strategy was not used further.

Finally, we tested *iteratively refining the prompt using model reasoning*, requiring two sentences of model reasoning as a JSON output field. Model reasoning helped identify why the model mislabeled certain materials (e.g., mistaking perimeter walls or foreground fences for siding), allowing us to iteratively add explicit instructions to focus on siding rather than architectural details and to select the primary material in mixed cases. This yielded a 4 percentage point accuracy increase over the initial prompt and was adopted as the final approach based on cost-accuracy tradeoffs.

Masonry Chimney

Seismic risk from a chimney depends on three components: presence, location (interior or exterior), and material (masonry or other). All three are important for correctly identifying risk. Figure 4(a) shows an example of an exterior chimney, and Figures 4(b) and 4(g) show interior chimney examples.

Two main strategies were evaluated for the masonry chimney prompt: *sequential questions to guide model behavior* and *explicit uncertainty accommodation*. Because three components are involved, each strategy could be applied in multiple ways. Presence, location, and material could be captured in one, two, or three sequential questions, and uncertainty could be allowed in any subset of the three fields (*presence* uncertainty — unclear if a chimney exists; *location* uncertainty — unclear if interior or exterior; *material* uncertainty — unclear if the material is masonry).

We evaluated combinations of one to three sequential questions with uncertainty in various fields; accuracy and false negative rate for each are shown in Figure 8. Allowing uncertainty in chimney presence decreased accuracy, whereas incorporating uncertainty in location reduced the false negative rate. The selected prompt, which was chosen for its balance of higher accuracy and lower rate of false negatives, incorporates two sequential questions and accommodates uncertainty in both location and material.

The selected prompt was further refined by assessing *patterns in uncertain and incorrect predictions*, which revealed a recurring issue: masonry chimneys painted to match the house’s siding (e.g., painted brick) were frequently misclassified as stucco. Adding an explicit note about this in the prompt improved accuracy.

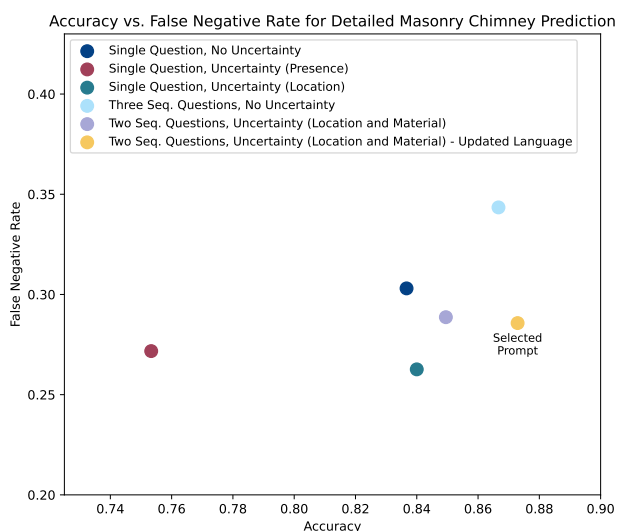


Fig. 8. Accuracy vs false negative rate for masonry chimney prompting strategies. Selected model incorporates sequential questions and uncertainty in both the location and material of the chimney.

Despite these improvements, the accuracy remains at or below 0.9 and the false negative rate above 0.25 across all prompts tested. Due to the high seismic risk that older masonry chimneys can pose, this performance is not satisfactory.

This likely reflects inherent challenges with street-level imagery: chimneys often sit at the rear of a house, partially or fully hidden from the street, and in densely spaced housing, camera angles can make it unclear which house a visible chimney belongs to. These limitations suggest that street-level imagery alone may be insufficient for this feature, and satellite or oblique imagery could be explored in future work.

Year of Construction

For Hayward, year of construction is available from tax parcel data for each building footprint [58]. In downstream seismic risk modeling, year of construction is used to map to categorical eras, which capture differences in seismic performance due to building code or construction practice changes. Era informs vulnerability model selection. Two downstream vulnerability models are relevant here: the commonly used Hazus fragility models, which differentiate eras at 1940 and 1975 [7], and the more detailed PEER-CEA models, which use 1945, 1956, and 1970 [9]. Year-based retrofit recommendation cutoffs are also important, including raised-foundation houses built before 1980 and houses with living space over garage, built before 2000. Thus, accurately predicting year of construction is important for both vulnerability model selection and retrofit eligibility.

For this task, we evaluated two prompt strategies: numerical predictions of year of construction (e.g., 1953) and categorical era-based predictions (e.g., ‘1940 to 1975’ for the Hazus model). Based on prior work using a similar categorical approach [13], we hypothesized that categorical prediction would perform better. However, the results showed that predicting the numeric year reproduced the year of construction from the local parcel data relatively well (Figure 9(a)). Mapping numeric year built to categories yielded higher accuracy than direct categorical prediction across all four categories (Figure 9(b)). Importantly, numeric prediction also more effectively captures some of the older buildings in the study area (Figure 9(c)) as opposed to categorical prediction (Figure 9(d)). These buildings are of particular interest given their higher expected seismic vulnerability.

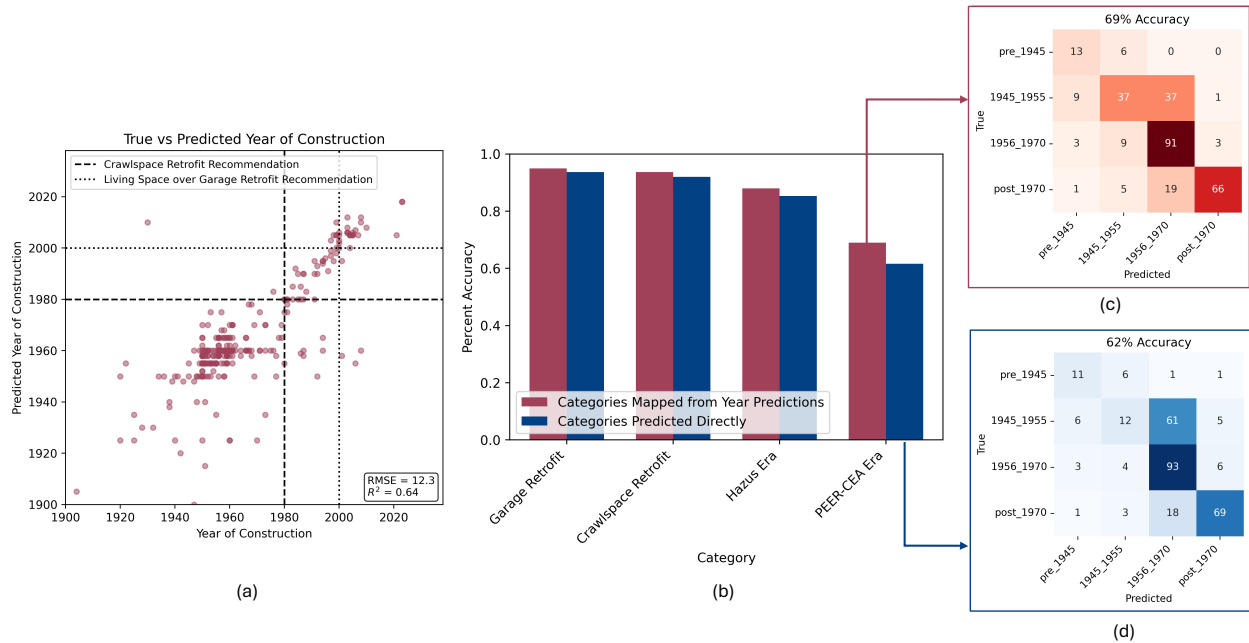


Fig. 9. Results for year of construction prompting strategies. (a) Prediction of numeric year of construction versus the true year of construction from the tax parcel data. (b) Percent accuracy for era classification across categories. Confusion matrix for PEER-CEA Classification using (c) numerical prediction and (d) categorical prediction.

Building Feature Overview

Final prompt accuracy across building features is summarized in Figure 10. Nearly all features meet or exceed 87% accuracy, with the majority well exceeding 90%. Two challenges persist in the final selected prompts: the masonry chimney’s high false negative rate and the lower accuracy for the PEER-CEA era assignment of older buildings. Future work will explore the direct use of these model outputs for regional seismic risk assessment, where the range of

acceptable error in vision model outputs will be better understood in context. Without going through this downstream modeling effort, we can only aim for the highest possible accuracy. Note that the accuracy values shown in Figure 10 come from a final run of the selected prompts and may differ slightly from those reported in the individual feature discussions above, as generative models do not produce identical outputs under identical prompting conditions.

Overall, results show that VLMs can effectively extract building features from street-level imagery with appropriate prompt design, though the value of specific prompt development strategies is not uniform across features. For example, *sequential questions* improve performance for some features, but not others. Compared to more traditional computer vision models, VLM prompting allows for the explicit inclusion of more detail, engineering context, and human reasoning, particularly benefiting ambiguous images and edge cases. Ultimately, the systematic application of these strategies provides a replicable framework for prompt development in similar building inventory tasks.

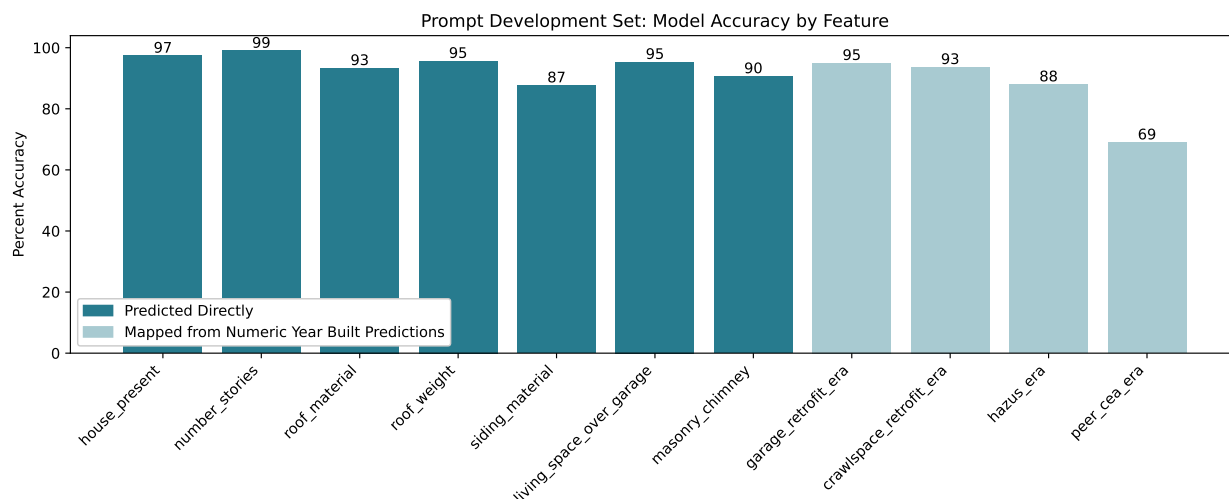


Fig. 10. Summary of Final Building Feature Prediction Accuracy for Prompt Development Image Set

PROMPT DEVELOPMENT FOR FOUNDATION DETAILS

The main goal of the foundation prompt is to determine the foundation type and, in the event of a crawlspace foundation, the height of the wood-frame first floor above the ground, as these features drive downstream vulnerability model selection. A secondary goal is to identify houses where the foundation type is not clear or difficult to distinguish between two possible types, enabling uncertainty propagation through regional seismic risk models. Throughout this section, various foundation prompts and postprocessing strategies are discussed, and all results are shown in Figure 13. Figure 13(a) shows the four accuracy metrics from Table 2, and Figure 13(b) shows recall for each foundation type.

Deterministic Foundation Outputs

The initial foundation detail prompt follows the setup described in Section 5. Similar to previous prompts, the initial foundation prompt requested a single ‘foundation type’ output, requiring the model to select between crawlspace, stem wall, and slab-on-grade foundations for each house. Requiring a single output selection is referred to as a deterministic foundation output. Several of the same prompt development strategies used for the building features prompt were used again to iterate on the initial foundation detail prompt.

As a first iteration, *few-shot prompting*, with examples of each foundation type, was evaluated; this did not improve performance, which points to the model struggling with ambiguity in the individual images, not necessarily a lack of understanding of the concept or foundation types themselves. This pattern was supported again by *evaluating patterns in uncertain and/or incorrect predictions*, as it seemed that the model was mostly able to classify clear cases, and was instead struggling with the more ambiguous ones. While minimal improvements in accuracy came from few-shot prompting, two of the previously used strategies did have large impacts on the foundation type predictions: *sequential questions* and *explicit uncertainty accommodation*.

In terms of using sequential questions, two main prompts were evaluated, one that directly asked about foundation type, containing descriptions of each foundation type and corresponding visual cues, and one with sequential questions,

which asked the model to return information on era, vent presence, entry stairs, lot slope, then finally the foundation type. We hypothesized that these questions would improve prediction quality by identifying relevant visual cues before the final classification. Additionally, as with roof and siding material, an ‘unclear’ label is added to the possible foundation type outputs. While an ‘unclear’ prediction is not helpful in informing risk or retrofit eligibility, it is preferred over an incorrect foundation type assignment.

The accuracy corresponding to each of these prompts is shown in Figure 13(a). For prompts that do not have an ‘unclear’ label, there is almost no impact of using sequential questions to guide model behavior (accuracy only changes by about 1 percentage point). On the other hand, in cases with an ‘unclear’ label included, sequential questions do impact performance; without sequential questions, the number of unclear foundations is very large, and with them, the quantity is greatly reduced. In spite of these differences, none of these prompt strategies is satisfactory for foundation prediction, with exact label accuracy around 50% and overall accuracy, which incorporates all four metrics discussed in Table 2, around 60%. Furthermore, as shown in Figure 13(b), these prompts also substantially overestimate the presence of slab-on-grade foundations and underestimate the presence of stem walls.

Probability-Based Foundation Outputs

Due to unsatisfactory performance, additional prompt development strategies beyond those used for screening and building features are required. As discussed in Section 4, there are instances in the ground truth data where houses are listed as having multiple possible foundation types. This arises in cases where there is insufficient information from the exterior view of the house to confidently assign a single foundation type. As a human labeler, one can typically rule out at least one foundation type relatively easily. For example, there can be ambiguity between a taller stem wall and a shorter crawlspace foundation (but clearly not slab-on-grade), or between a shorter stem wall and slab-on-grade (but clearly not a crawlspace). Thus, we shifted the prompt to accommodate this uncertainty by predicting foundation type probabilities (summing to one) rather than just assigning a single foundation type label. For example, instead of outputting {‘stem_wall’}, the model outputs {‘stem_wall’:0.7, ‘crawlspace’:0.05, ‘slab’:0.25}. This probability output is not the probability the model produces internally, but rather a generative output (verbalized probability). We hypothesized that outputting the probability of different foundation types may 1) improve model reasoning or 2) provide more transparency in the predictions, allowing the model to effectively rule out foundation types. The *sequential questions to guide model reasoning* were kept as part of the probability-based foundation prompt.

To evaluate whether probability-based foundation predictions improved model reasoning, we can compare the initial deterministic predictions (with sequential questions, without uncertainty) against the ‘most likely’ output: the foundation type with the highest weight from the probability-based assignment. As shown in the 2nd and 5th columns of Figures 13(a) and 13(b), the performance of these two models is very similar. Thus, probability-based outputs do not improve model reasoning relative to a model that returns a single foundation type label.

The second hypothesis is that the model transparency revealed by the probability outputs could help correctly designate the foundation type and/or the uncertainty in the predictions. To probe this, we move beyond just the most likely (highest weight) foundation type. If we take all non-zero probability values as possible foundation types, the majority of houses have all three possibilities. This is likely because the model interprets any visual ambiguity as grounds for assigning non-zero weight to all types. This is not a helpful output, as it indicates that almost all houses have several possible foundation types; thus, postprocessing is developed to refine model outputs.

The first step in processing the probability-based results is to enact a lower bound value for what is considered a possible foundation type: if one of the foundation types is predicted as having a probability of 0.1 or less, it is dropped from the possible foundation types for that house. The threshold value of 0.1 was determined based on a sensitivity study that minimized the number of houses retaining all three foundation types while preserving correct types in the possible set. Prior to the lower bound filter, the majority of houses had three foundation types (Figure 11(a)), and the filter effectively results in the large majority of houses now having two possible foundation types (Figure 11(b)). Based on the confusion matrix shown in Figure 11(b), the filter rarely eliminates a foundation type that the ground truth considers possible. For example, there are very few slab foundations that fall into the predicted ‘crawlspace or stem wall’ category, and there are very few crawlspace foundations that fall into the predicted ‘stem wall or slab’ category. This demonstrates that the model is almost always able to correctly rule out at least one foundation type, as very few houses remain with three possible foundation types in Figure 11(b).

Beyond the initial lower-bound filter, the majority of houses still have two possible predicted foundation types, either stem wall and slab, or stem wall and crawlspace. This is not surprising because a stem wall is visually intermediate compared to the other two foundation types, meaning that different stem walls can look similar to either crawlspace or slab-on-grade foundations. However, there are nine times more predicted cases with multiple possible foundation

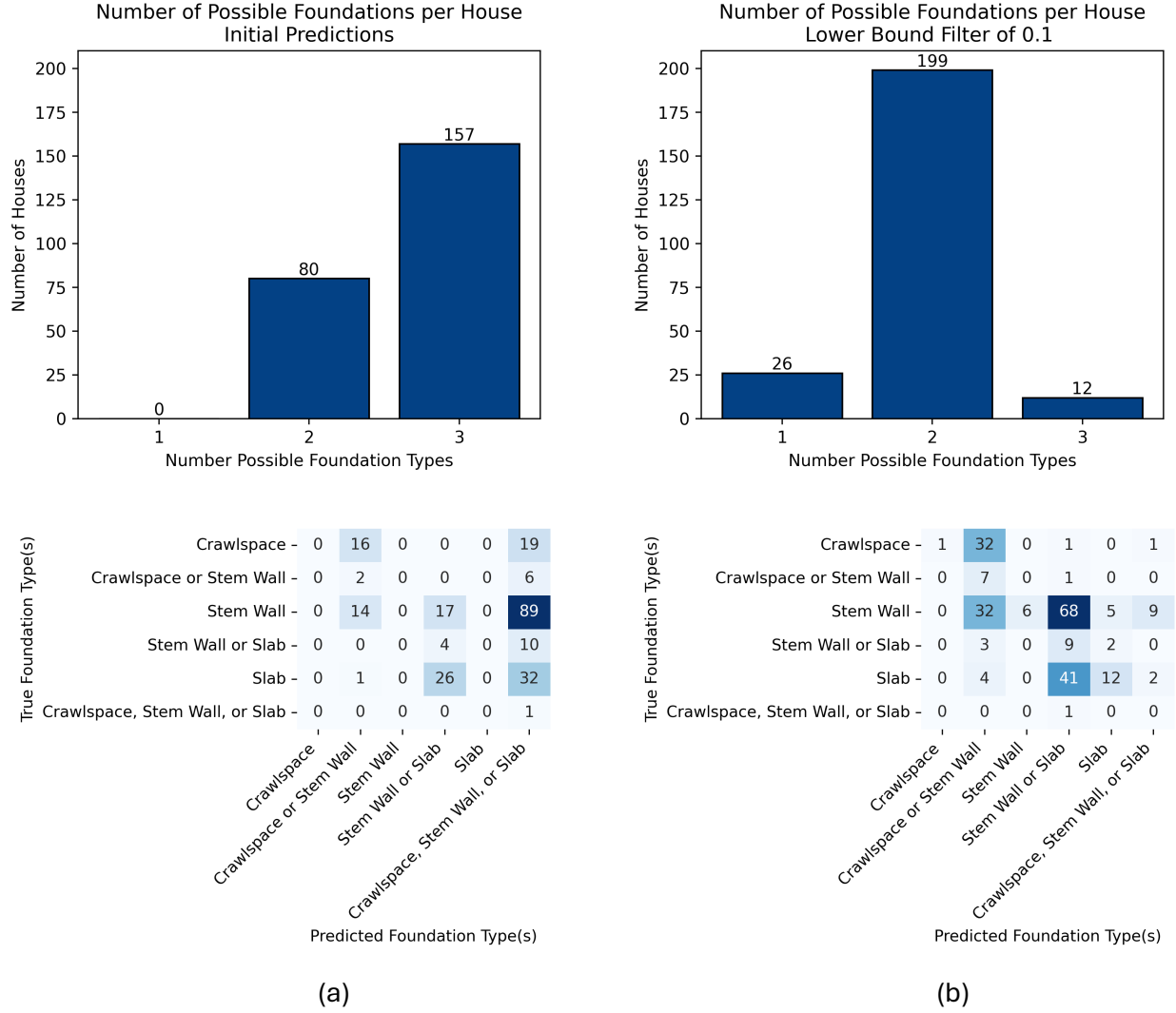


Fig. 11. Number of houses with 1, 2, or 3 possible foundation types and associated foundation prediction confusion matrix (a) without lower-bound probability filter and (b) with lower-bound probability filter.

types than there are in the ground truth data, so further postprocessing is developed to improve prediction specificity.

As a next step, one could use the probability values across foundation types and assign thresholds, similarly to the lower-bound threshold. However, VLMs do not necessarily produce calibrated probability values across different houses evaluated in separate API calls; for example, a stem wall predicted with 0.6 probability for one house does not necessarily reflect the same underlying confidence as a crawlspace predicted with 0.6 probability for another house. This is because the model cannot relate evidence across separate API calls to create a common baseline, and evaluating all houses within a single call to enforce such a baseline would be infeasible given the token count.

Thus, there are two continuous-variable outputs that could be used to further specify possible foundation types for houses with multiple remaining foundation types. First is the difference between the predicted probability values. The relative difference between predicted probabilities is more robust to cross-building variation in model confidence, since it captures only the internal ranking of foundation types within a single model response, rather than the absolute magnitude of any individual probability. Filtering possible foundation types based on the difference in predicted probability values is referred to as the ‘probability filter.’ Second is the predicted foundation height (in inches). The three foundation types differ systematically in how much they elevate the structure above the ground; slab-on-grade foundations sit at or near ground level, stem walls elevate the structure by a modest amount, and crawlspace foundations

require sufficient height to construct small wood-frame walls. Because these height differences are visually detectable from the building exterior, predicted foundation height provides a physically grounded basis for filtering possible foundation types, referred to as the ‘height filter.’

To develop the probability and height filters, threshold values are needed to differentiate between 1) ‘stem wall,’ ‘crawl space,’ and ‘stem wall or crawl space’ foundations and 2) ‘stem wall,’ ‘slab,’ and ‘stem wall or slab’ foundations. Thresholds were determined by plotting height and difference in probability values for each foundation type present in the prompt development set and evaluating whether there were meaningful differences in these metrics across foundation types based on the ground truth data. Threshold values were selected to maximize specificity while minimizing incorrect predictions in the prompt development set.

The foundation accuracy metrics and recall values are shown in Figures 13(a) and 13(b) for cases using 1) only the lower-bound filter, 2) lower-bound and probability filters, and 3) lower-bound and height filters. Using only the lower-bound filter, the overall accuracy is higher than other models, but the exact label accuracy value is low. Incorporating either the probability or height filters in postprocessing improves the exact label accuracy by 19-24 percentage points and reduces cases with multiple possible foundation types, at the cost of 5-8 percentage points more fully incorrect predictions. Neither filter reliably differentiates stem wall and slab foundations, and there are a substantial number of remaining cases that have both of these as possible foundation types. However, the height filter does a better job of differentiating between stem wall and crawl space foundations, which is important when quantifying expected seismic performance, and results in higher accuracy metrics than the probability filters; thus, this postprocessing strategy is selected for future use. Using the height filter to further rule out possible foundation types improves performance and allows us to leverage the model transparency enabled by probability-based outputs. Thus, the lower-bound filter and height filter make up the selected postprocessing to assign foundation types.

Neighboring Images to Inform Predictions

Based on the recall values, shown in Figure 13(b), the lower-bound and height filter postprocessing strategy gives us a relatively robust way to identify the crawl space foundations; however, it is still lacking in terms of differentiating between stem wall and slab foundations. In the United States, particularly following World War II, neighborhoods are often developed by a single developer, which can lead to construction and design similarities across neighboring houses. The advent of this style of construction aligns with the increasing use of slab-on-grade foundations in the 1960s. Thus, we hypothesize that leveraging information on neighboring houses may improve foundation classification, particularly for the stem wall / slab differentiation.

There are several ways that information on neighboring houses could be included in a prompt; in this study, we evaluate the effectiveness of running a second pass of the foundation prompt. The first pass is as described above; then, the second pass provides the initial height and probability-based foundation type predictions for a given house, alongside one street-level image of each of the four neighboring houses. Neighboring houses are selected by finding the two closest footprints to the original house, then the closest footprint to each of those two, so as to identify houses along the same side of the street in typical street configurations. Imagery is obtained using the same methods described in Section 4, each image is screened using the screening prompt developed in Section 7, and then the single closest screened image to each neighboring house is adopted for use. Using the original answers as a starting point, the model is instructed to compare the target house to the neighboring houses and consider whether any foundation characteristics might be inferred more reliably given similarities or differences across the neighborhood. The prompt specifies that just because one house has a certain feature, it does not mean the target house has it, but if multiple neighbors share a feature that is also somewhat visible on the target house, that may increase confidence. The model is instructed to change its prediction of the foundation probability values and the foundation height only if the additional neighborhood context increases confidence that the original values were incorrect.

Two tests were used to evaluate the efficacy of the additional second-pass prompt using images of neighboring houses. The first case used the second-pass prompt for all houses, regardless of the original prediction. The second case used the second-pass prompt only for cases that still had multiple possible foundation types after the first foundation prompt. Overall model accuracy increased when the second-pass neighbor evaluation was included, particularly the exact label accuracy, which increased by 10 percentage points, as shown in dark blue in Figure 13(a). Most notably, the inclusion of the neighboring images greatly improved the recall of the slab-on-grade foundations (Figure 13(b)). Overall recall remained high, and the percentage of slab-on-grade houses identified with only the correct label increased from 20% to 56%. The two tests (rerunning all houses versus just those with multiple foundation types after the first pass) performed similarly, but the latter significantly reduced the number of API calls required, and thus was selected.

To illustrate the impacts of the second-pass foundation prompt, we ran the two-pass prompt on an example

neighborhood. Figure 12(a) shows the foundation type predictions for the example neighborhood given only images of the target house, with approximately 38% of predictions having two or more foundation types. With the second pass, shown in Figure 12(b), there is increased neighborhood consistency, particularly for the slab-on-grade foundations, and the number of houses with two or more foundation types is reduced by 15 percentage points.

The increase in accuracy and recall on the prompt development set represents a real benefit of including neighboring images relative to ground truth; however, this second-pass prompt likely also introduces some speculation. In cases where the ground truth is not visible, the resulting predictions may be more cohesive across the neighborhood, but this does not guarantee that these predictions are correct.

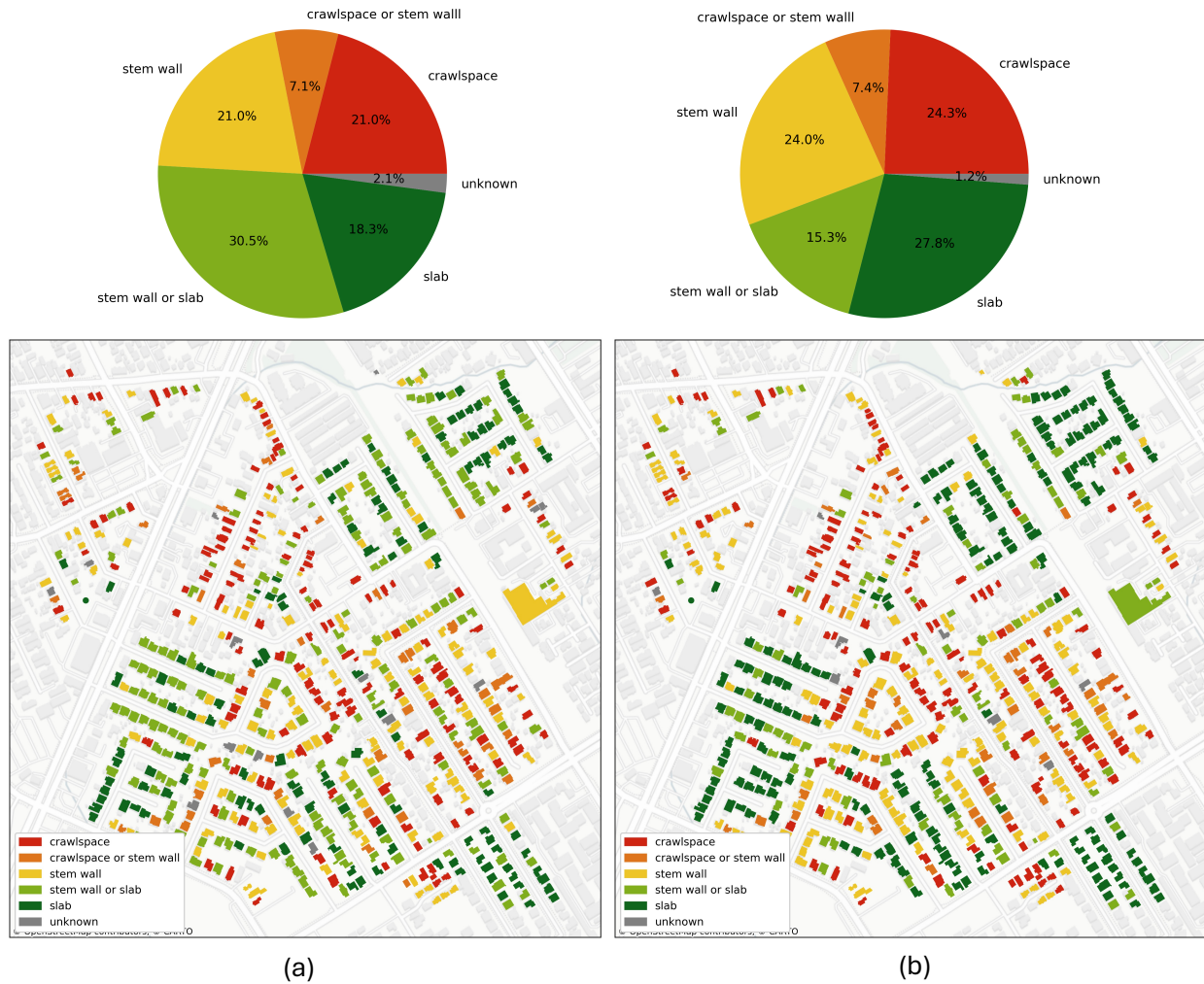


Fig. 12. Impact of second-pass prompt that reassesses foundation type, given images of nearest four neighboring houses for example neighborhood. (a) Initial predictions and (b) updated predictions of foundation type(s) by house and proportion of foundations across an example neighborhood.

Prompting Multiple Times and Taking Majority Vote

The final strategy that is evaluated for the foundation prompt stems from the hypothesis that the ambiguity in the foundation type would be partially resolved using multiple passes of the same prompt over the same images. For example, one could prompt the same way three times and assess the combined results, potentially obtaining a more accurate prediction. We evaluated this strategy by rerunning the foundation prediction, height filter, and neighborhood predictions three times and taking the majority vote label. However, this greatly increased computational time and cost, without providing a corresponding benefit, as shown in Figure 13(a) and 13(b), and therefore is not adopted.

Foundation Prediction Overview and Comparison with Humans

In summary, there are many ways to evaluate the foundation type, including models with single-label outputs and models with probability-based outputs. Based on overall results, there are clear benefits of using probability-based outputs, which allow for improved ability to rule out one or more foundation types. There are also benefits in conducting a second pass with neighboring images, as it increases accuracy and substantially improves the recall of slab-on-grade foundations. Figure 13 summarizes the accuracy and recall across the foundation prompting strategies evaluated in this study. The model that is selected for future use incorporates 1) probability-based outputs, 2) postprocessing based on predicted foundation height, and 3) a second-pass foundation prompt using neighboring images for cases originally predicted as having multiple possible types.

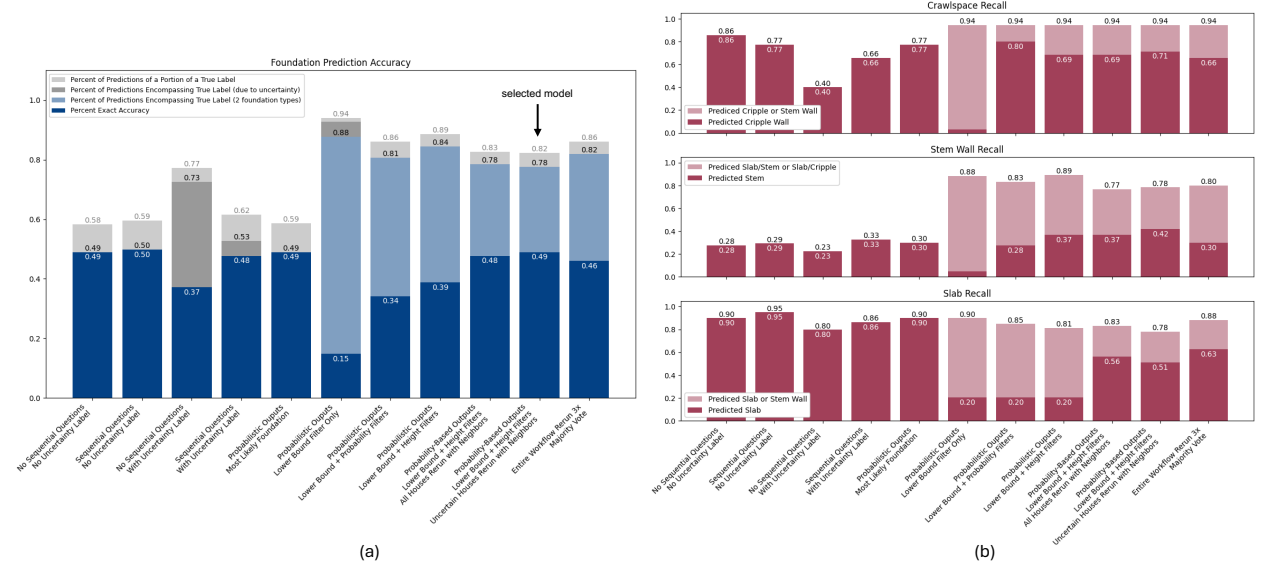


Fig. 13. Foundation prediction accuracy across deterministic and probabilistic outputs. (a) Overall prediction accuracy across prompts. (b) Recall for crawlspace, stem wall, and slab-on-grade foundations across prompts.

As an assessment of model performance, we compare it against the performance of a human performing the same task. A group of ten structural engineering PhD students, postdocs, and staff was recruited to label foundation types for the prompt development set. Participants could select a single foundation type or, when uncertain, two to three foundation types for a given house. In other studies related to natural hazard risk, graduate and undergraduate students are employed in image-tagging efforts [69]. While not specialists in residential foundation types, participants' structural engineering backgrounds make them representative of labelers typically employed in similar image-tagging efforts.

As shown in Figure 14, the selected model achieves higher exact label accuracy than most individual human labelers, and it meets or exceeds the overall accuracy of all ten individuals. Two aggregate human metrics were also computed for comparison: the average accuracy across all labelers, and the accuracy of the majority vote label, defined as the most frequently selected foundation type(s) across the ten participants. The selected foundation model outperforms the human average across all four accuracy metrics (Table 2). Compared to the human majority vote, the selected model does not achieve as high an exact label accuracy. However, it matches the human majority vote for overall accuracy. It is also worth noting that the majority vote label represents consensus across ten labelers, representing a standard that would rarely be achievable in practice. On average in the labeled data, only approximately half of participants selected the majority answer. Since labeling efforts generally rely on a single individual, the selected model appears to offer greater reliability than a single human label.

It is also possible that the model could be used as an assistive tool to humans, where the model could be used as a first-pass evaluation, and then human experts could review the uncertain cases. Assuming a human expert could identify the foundation type with 100% accuracy (unlike the human labelers shown in Figure 14), the resulting exact label accuracy would be 79% and overall accuracy would be 83%, while reducing the human workload by 66%.

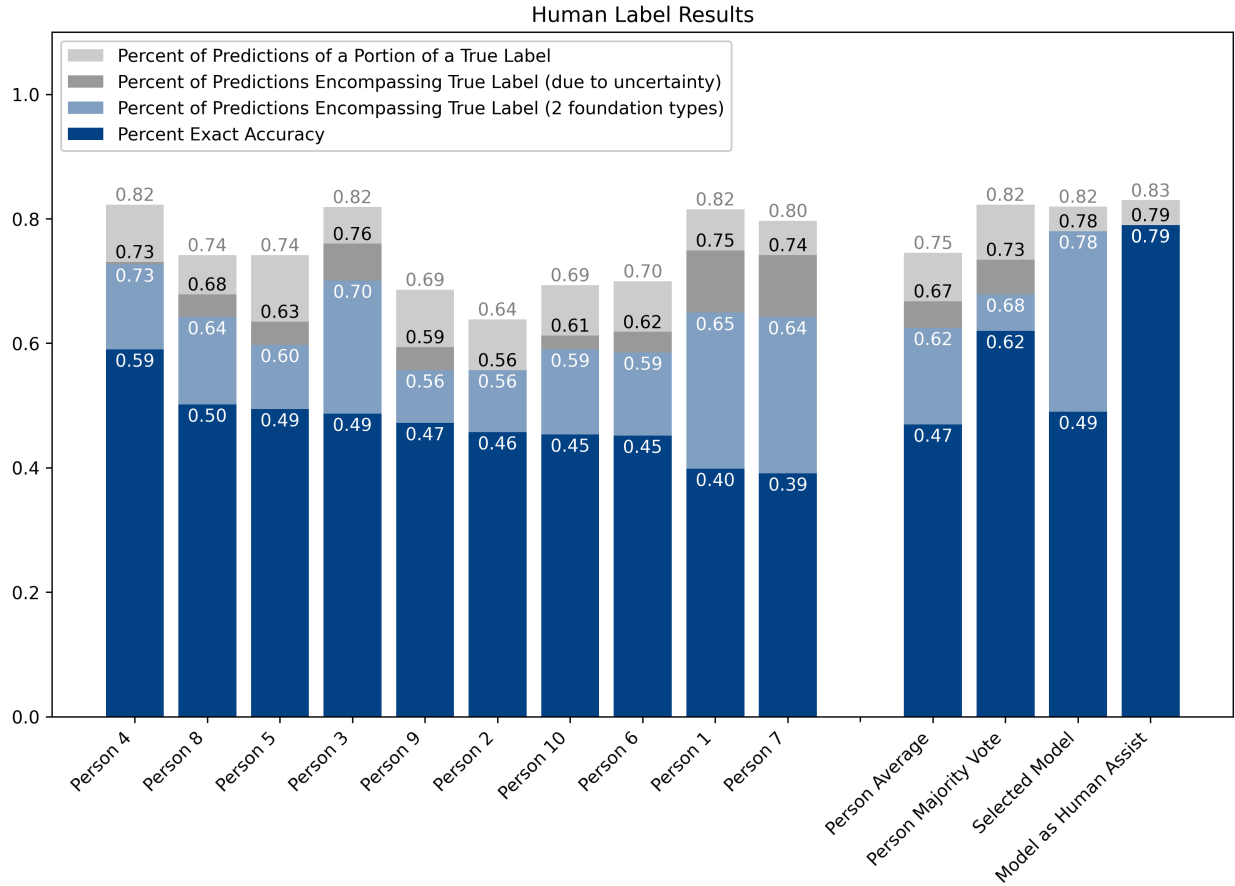


Fig. 14. Individual human labeler accuracy, average human accuracy and majority vote accuracy compared to selected foundation model.

MODEL EVALUATION AND GENERALIZATION

The prompt development image set, which consists of street-level images of Hayward, is used to develop and select final versions of the screening, building feature, and foundation detail prompts. All prior results are reported on the prompt development set.

To test model performance and evaluate generalizability across locations, the selected screening, building feature, and foundation detail prompts are evaluated using the four image test sets, drawn from four different sites across California (see Section 4). Only sites in California are used because the retrofit programs and structural models are developed in the context of California and are not widely generalizable elsewhere.

Figure 15 demonstrates that the screening and building feature models are robust across the four test cities, with predictable and acceptable levels of accuracy for building inventory development. Compared to the prompt development set, the screening prompt has accuracy within 2 percentage points, and most building feature tasks have accuracy within 7 percentage points. One exception is the ambiguity in the number of stories for older houses in Los Angeles and Oakland due to small attic rooms and partial upper stories, which leads to lower accuracy in classifying the number of stories for those test cases, shown in Figure 4(e) and 4(f). In addition, Los Angeles has a higher accuracy for masonry chimney prediction, but this mainly arises because there are fewer masonry chimneys in the sampled neighborhood of Los Angeles overall. Similarly, there are cases of water coolers or air conditioners on the roofs of houses that are predicted as masonry chimneys; this is not observed in the prompt development set, and therefore causes a reduction in accuracy for chimney predictions in Oakland.

Based on the four test sites, the foundation prediction is also robust across test locations. As shown in Figure 16, there is no noticeable drop in accuracy between the prompt development set and the four test sets, implying strong

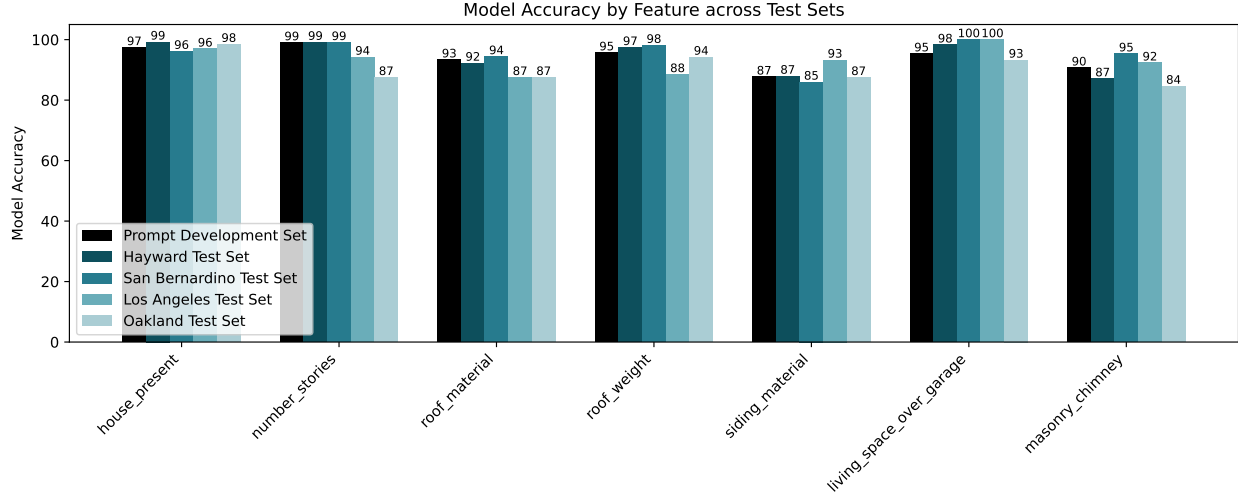


Fig. 15. Screening and building feature prediction accuracy across four test sites.

model generalizability. The four accuracy metrics from Table 2 are shown for the prompt development set and four test sets in Figure 16(a). Across the four test locations, overall accuracy remains within 5 percentage points of that for the prompt development image set. More specifically, the breakdown of accuracy values, including exact accuracy, is similar for the prompt development set, Hayward test set, and San Bernardino test set because these image sets have a similar distribution of foundation types. The Los Angeles and Oakland test sets have higher exact accuracy values because they have a higher proportion of crawlspace foundations; furthermore, the neighborhoods selected for those two test sets are older parts of these cities, and in such houses, crawlspaces can be taller and more visually obvious. The model is selected to be sensitive to crawlspace foundations (due to their seismic vulnerability), so it is unsurprising that there is higher exact accuracy for case study cities with a higher proportion of this foundation type. Figure 16(b) shows the recall for the crawlspace foundations, which demonstrate patterns similar to the overall foundation prediction accuracy. Across Figures 16(a) and 16(b), values are stabilized by controlling for the year of construction and lot slope, which can both block the use of a standard seismic retrofit design.

IMPLEMENTATION DETAILS

Initially, this study explored the use of both open-source and proprietary models, including Qwen and OpenAI models. Based on an initial evaluation, the Qwen models (Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct [70, 71]) did not perform well on the majority of building feature tasks. Thus, most evaluation was done using two primary OpenAI models: the GPT-5-2025-08-07 and GPT-5.1-2025-11-13 snapshot models. Snapshot models were used to improve consistency and standardization of model behavior over the course of experimentation. Ultimately, GPT-5-2025-08-07 was selected for use, and all figures and reported results in this study are based on predictions using this snapshot model. GPU hardware details were not exposed by the hosted API provider.

All prediction was performed through the hosted OpenAI Responses API. Structured outputs were used to return JSON-formatted responses with standardized field names and data types. For the screening and building feature prompts, all inference was performed using the `image_detail = "low"` setting, as this selection did not noticeably affect performance and reduced cost. For the foundation detail prompt, inference was performed using `image_detail = "high"`, which improved performance for this task. All inference was done using medium reasoning effort, which seemed to balance performance and cost based on initial testing.

The cost of the screening, building feature, and foundation detail prompts was approximately \$12 for the prompt development image set (345 buildings and ~1,400 images) and approximately \$5-\$6 for each test image set (130 buildings and ~550 images each). As part of ongoing work by the authors, the three prompts were subsequently deployed across the case study city of Hayward. The OpenAI Batches API was used for the city-wide deployment, which is an asynchronous processing mode that accepts large jobs and returns results with a delay of up to 24 hours (though in practice much shorter). This API offers a cost discount relative to the synchronous Responses API, and performance was consistent based on a spot-check of outputs across both APIs. Evaluating approximately 27,700

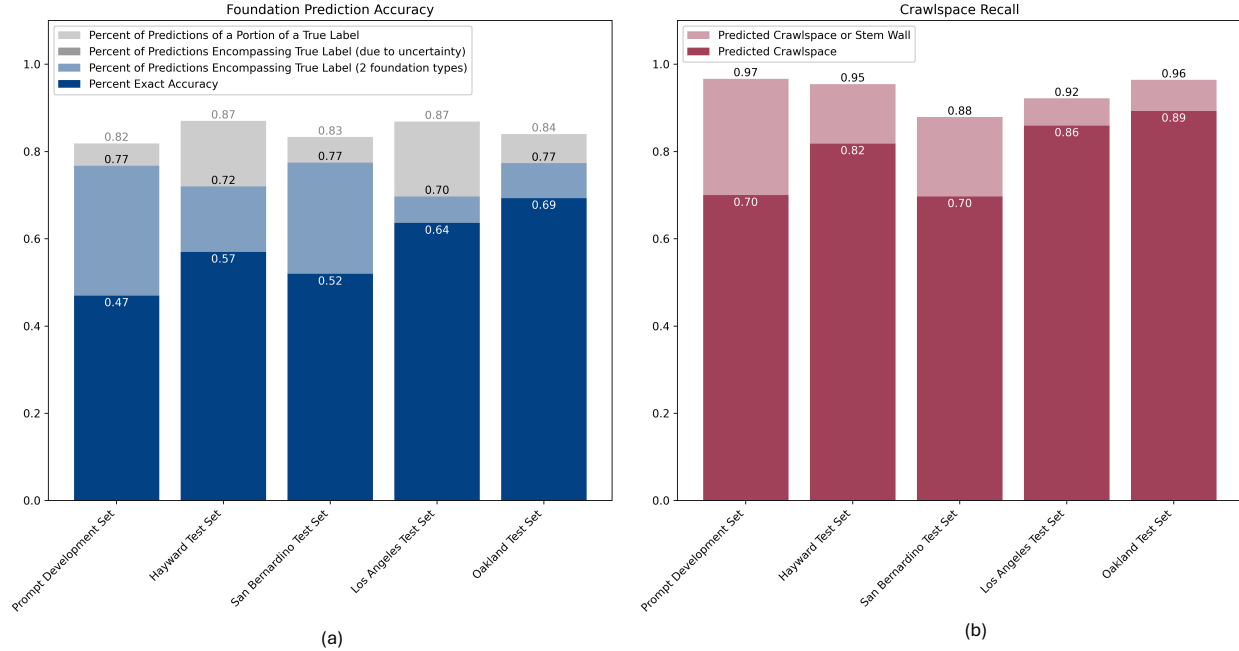


Fig. 16. Model performance across test sites for (a) overall accuracy and (b) recall for crawlspace foundations.

residential buildings (~95,000 images) using the three prompts developed in this study and the OpenAI Batches API cost \$613 in total. Evaluation of imagery to collect building inventory information at this scale is often viewed as cost- and time-prohibitive for humans, who instead would typically sample a subset of buildings and extrapolate to estimate city-wide risk, rather than obtaining building-specific data for the full inventory. Thus, this workflow enables building-level data collection relevant for seismic risk assessment that is not readily available in existing datasets or feasible via manual data collection methods at comparable scale, cost, and accuracy.

LIMITATIONS AND FUTURE WORK

A feature of this study that poses both a benefit and a limitation for downstream risk assessment is the presence of uncertainty in the predicted foundation types (among other features). An uncertain foundation type in the model response is preferred over an incorrect foundation type prediction, and the flexibility of VLMs enables these types of outputs, which is beneficial. However, propagating uncertainty in the foundation type leads to corresponding uncertainty in the quantified seismic risk in a community. Future work could aim to increase accuracy and reduce uncertainty in predicted foundation type, though the extent to which this is possible using only exterior building views is somewhat limited. Furthermore, it is impossible to tell the retrofit status of raised foundations from exterior views, so incorporating that information into a regional risk assessment study would require additional information beyond imagery, such as building permit data.

In addition, street-level coverage is not uniform across neighborhoods, and some houses may have no usable imagery, limiting the completeness of any inventory derived from this approach. This gap may be spatially correlated with building age or neighborhood characteristics relevant to seismic risk. Furthermore, the building typologies and prompts are developed and evaluated using California construction typologies and retrofits, and generalizability to other regions would require additional evaluation.

VLMs are rapidly evolving, and the models used in this study represent one point in that evolution. While the exact prompting and performance are tied to the model used in this study, we present a number of repeatable strategies for developing building inventory prompts, which can be applied in other settings and to future models. We also present accuracy measures that can be useful if both the underlying ground truth label and the predicted label may contain uncertainty (e.g., foundation type). While the method is applied here for the specific task of gathering seismic risk information, the same strategy can also be used for other downstream tasks, such as extracting relevant building features for hurricane risk (e.g., roof shape, eaves, window-to-wall ratio, etc.), flood risk (e.g., first-floor elevation), and urban heat impacts (e.g., roof and siding material and color, vegetation, etc.), or other visible building attributes.

Finally, this study leverages similarities in tract-style development in the United States as a tool for improving predictions. Future work could further evaluate patterns across neighborhoods, correlations between building features, and correlations based on proximity to inform data imputation and inventory development.

CONCLUSIONS

This study assesses VLMs for the purpose of identifying building features to inform seismic vulnerability in residential construction from street-level imagery. This study uses three individual prompts to 1) screen, 2) identify building features, and 3) assess foundation details based on street-level images of buildings in California.

Overall, zero-shot prompts can screen for images showing single-family homes and detect building features with a high degree of accuracy. However, this accuracy is partially contingent on robust prompts. Prompts can be evaluated and improved using the set of prompt development strategies identified and used in this study. The screening and building feature prompts perform similarly across the four California test sites, demonstrating the generalizability of these models across single-family residential construction in California.

Foundation type prediction is more challenging due to occlusions in the imagery and inherent limitations in determining foundation type using only exterior building views. Two additional strategies are effective in improving prediction accuracy, including probability-based outputs and a second-pass prompt for uncertain foundation types, which uses images of neighboring houses to improve accuracy and specificity, particularly for slab-on-grade foundations. Even so, prediction accuracy remains limited, and for that reason, it is critical to acknowledge and propagate uncertainty in model predictions. In the context of a regional seismic risk assessment, it is better to acknowledge uncertainty than make a confident incorrect prediction, as uncertainty can be propagated through the risk assessment, whereas incorrect predictions can cause bias.

Although foundation type models have limited accuracy even with all prompt development strategies and uncertainty propagation, they match or exceed the accuracy of human labelers and perform robustly across test locations. Critically, foundation type is typically unavailable from existing inventory data sources, so even uncertain predictions offer more detail than is currently available.

Overall, this study demonstrates that with systematic prompt development and evaluation, zero-shot prompting of VLMs can be used to effectively extract building information from street-level imagery. Together, these models provide a reproducible framework for collecting detailed building inventory data to support regional seismic risk assessments.

Software and Data Availability

All prompts and postprocessing scripts used in this study, along with building footprints and the corresponding ground-truth labels, are available on GitHub at <https://github.com/mlochhead/residential-seismic-vlm>. Street-level imagery is not redistributed but can be retrieved using the included download scripts and footprint geometries.

Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program, a Stanford Graduate Fellowship, and the Blume Center for Earthquake Engineering. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

- [1] FEMA. *FEMA P-1100, Vulnerability-Based Seismic Assessment and Retrofit of One- and Two-Family Dwellings: Volume 1 - Prestandard*. Federal Emergency Management Agency, 2019.
- [2] U.S. Census Bureau. American Community Survey, 2022.
- [3] JF Hall. Wood Buildings. *Earthquake Spectra*, 12:125–176, 1996.
- [4] Mary C Comerio. *Northridge Housing Losses: A Study for the California Governor's Office of Emergency Services*. Center for Environmental Design Research, University of California at Berkeley, 1995.
- [5] Pablo Heresi and Eduardo Miranda. Fragility curves and methodology for estimating postearthquake occupancy of wood-frame single-family houses on a regional scale. *Journal of Structural Engineering*, 147(5), 2021.
- [6] SPUR Shelter in Place Task Force. *Safe Enough to Stay*, 2012.
- [7] FEMA. Hazus 6.1 Earthquake Model Technical Manual. Technical report, Federal Emergency Management Agency, 2024.
- [8] CEA. Understanding earthquake structural risks by house type. <https://www.earthquakeauthority.com/strengthen-your-house>, 2026. Accessed: August 2026.
- [9] D. Welch and G. Deierlein. Technical background report for structural analysis and performance assessment. Technical Report PEER Report Number 2020/22, Pacific Earthquake Engineering Research Center, 2020.
- [10] U.S. Army Corps of Engineers. National Structure Inventory. <https://www.hec.usace.army.mil/confluence/nsi/technicalreferences/latest/technical-documentation>, 2022.
- [11] FEMA. Hazus 7.0 Inventory Technical Manual. Technical report, Federal Emergency Management Agency, 2025.
- [12] Laxman Dahal, Henry Burton, and Kuanshi Zhong. High-fidelity high-resolution regional seismic risk and resilience assessment of large building inventories. *Earthquake Engineering & Structural Dynamics*, 54(5): 1376–1396, 2025.
- [13] Zichao Zeng, June Moh Goo, Xinglei Wang, Bin Chi, Meihui Wang, and Jan Boehm. Zero-shot building age classification from facade image using gpt-4. *arXiv preprint arXiv:2404.09921*, 2024.
- [14] Fei Pan, Sangryul Jeon, Brian Wang, Frank McKenna, and Stella X Yu. Zero-shot building attribute extraction from large-scale vision and language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8647–8656, 2024.
- [15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [16] Daniela Gonzalez, Diego Rueda-Plata, Ana B Acevedo, Juan C Duque, Raúl Ramos-Pollán, Alejandro Betancourt, and Sebastian García. Automatic detection of building typology using deep learning methods on street level images. *Building and Environment*, 177, 2020.
- [17] Federica Ghione, Steffen Mæland, Abdelghani Meslem, and Volker Oye. Building stock classification using machine learning: A case study for oslo, norway. *Frontiers in Earth Science*, 10, 2022.
- [18] Patrick Aravena Pelizari, Christian Geiß, Paula Aguirre, Hernán Santa María, Yvonne Merino Peña, and Hannes Taubenböck. Automated building characterization for seismic risk assessment using street-level imagery and deep learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 180, 2021.
- [19] Hafidz R Firmansyah, Prasanti Widyasih Sarli, Andru Putra Twinanda, Devin Santoso, and Iswandi Imran. Building typology classification using convolutional neural networks utilizing multiple ground-level image process for city-scale rapid seismic vulnerability assessment. *Engineering Applications of Artificial Intelligence*, 131: 107824, 2024.

- [20] Rony Kalfarisi, Maadh Hmosze, and Zheng Yi Wu. Detecting and geolocating city-scale soft-story buildings by deep machine learning for urban seismic resilience. *Natural Hazards Review*, 23(1), 2022.
- [21] Qian Yu, Chaofeng Wang, Frank McKenna, Stella X Yu, Ertugrul Taciroglu, Barbaros Cetiner, and Kincho H Law. Rapid visual screening of soft-story buildings from street view images using deep learning classification. *Earthquake Engineering and Engineering Vibration*, 19:827–838, 2020.
- [22] Chaofeng Wang, Sarah Elizabeth Antos, and Luis Miguel Triveno. Automatic detection of unreinforced masonry buildings from street view images using deep learning-based image segmentation. *Automation in Construction*, 132, 2021.
- [23] Peng-Yu Chen, Zheng Yi Wu, and Ertugrul Taciroglu. Classification of soft-story buildings using deep learning with density features extracted from 3d point clouds. *Journal of Computing in Civil Engineering*, 35(3), 2021.
- [24] Chaofeng Wang, Sascha Hornauer, Stella X Yu, Frank McKenna, and Kincho H Law. Instance segmentation of soft-story buildings from street-view images with semiautomatic annotation. *Earthquake Engineering & Structural Dynamics*, 52(8), 2023.
- [25] Jian Kang, Marco Körner, Yuanyuan Wang, Hannes Taubenböck, and Xiao Xiang Zhu. Building instance classification using street view images. *ISPRS journal of photogrammetry and remote sensing*, 145:44–59, 2018.
- [26] Deepika Raghu, Martin Juan José Bucher, and Catherine De Wolf. Towards a ‘resource cadastre’ for a circular economy–urban-scale building material detection using street view imagery and computer vision. *Resources, Conservation and Recycling*, 198, 2023.
- [27] Yan Li, Yiqun Chen, Abbas Rajabifard, Kourosh Khoshelham, and Mitko Aleksandrov. Estimating building age from google street view images using deep learning. In *10th International Conference on Geographic Information Science*, 2018.
- [28] Matthias Zeppelzauer, Miroslav Despotovic, Muntaha Sakeena, David Koch, and Mario Döller. Automatic prediction of building age from photographs. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 126–134, 2018.
- [29] Miguel Ureña-Pliego, Rubén Martínez-Marín, Beatriz González-Rodrigo, and Miguel Marchamalo-Sacristán. Automatic building height estimation: Machine learning models for urban image analysis. *Applied Sciences*, 13(8):5037, 2023.
- [30] Zhen Xu, Furong Zhang, Yingying Wu, Yajun Yang, and Yuan Wu. Building height calculation for an urban area based on street view images and deep learning. *Computer-Aided Civil and Infrastructure Engineering*, 38(7): 892–906, 2023.
- [31] Vitor Silva, Romain Sousa, Feliz Ribeiro Gouveia, Jorge Lopes, and Maria Joao Guerreiro. A building imagery database for the calibration of machine learning algorithms. *Earthquake Spectra*, 40(2), 2024.
- [32] Rajesh Kumar, Claudio Rota, Flavio Piccoli, Gianluigi Ciocca, et al. Deep learning for building attribute classification from street-view images for seismic exposure modeling. *Applied Sciences*, 16(2), 2026.
- [33] Shayan Shourabi and Ali Bakhshi. Rapid visual screening of buildings for potential seismic hazards: automated deep-learning classification approach. *Journal of Computing in Civil Engineering*, 39(3):04025016, 2025.
- [34] Barbaros Cetiner, Frank McKenna, Sang-Ri Yi, Brian Wang, and Ioannis Vouvakis Manousakis. Brails++: Version v4.1.3, 2025. URL <https://doi.org/10.5281/zenodo.16910168>.
- [35] Adam Zsarnóczy, Gregory G. Deierlein, Frank McKenna, Matthew Schoettler, Sang-Ri Yi, Barbaros Cetiner, Aakash Bangalore Satish, Jinyan Zhao, Justin Bonus, Abiy F. Melaku, Sina Naeimi, Pedro Arduino, Rachel Davidson, Catherine Gorle, Sanjay Govindjee, Ahsan Kareem, Tracy L. Kijewski-Correa, Laura N. Lowes, Michael Motley, Seymour M. J. Spence, Ertugrul Taciroglu, Alexandros A. Taflanidis, and Matthew DeJong. An open-source simulation platform to support and foster research collaboration in natural hazards engineering. *Frontiers in Built Environment*, 11, 2025. doi: 10.3389/fbuil.2025.1590479. URL <https://doi.org/10.3389/fbuil.2025.1590479>.

- [36] Ziyu Peng, Weisheng Lu, Hongda An, Xianhua Xia, Yi Zhang, Fan Xue, and Junjie Chen. Vision language model (vlm)-enabled street view analytics: a systematic literature review. *Engineering, Construction and Architectural Management*, pages 1–19, 2025.
- [37] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pages 3992–4003. IEEE, 2023.
- [38] Siyuan Yao, Siavash Ghorbany, Matthew Sisk, Ming Hu, and Chaoli Wang. Leveraging zero-shot learning on street-view imagery for built environment variable analysis. In *International Symposium on Visual Computing*, pages 243–254. Springer, 2024.
- [39] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024.
- [40] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.
- [41] Yuandong Pan, Linjun Lu, Mudan Wang, Varun Kumar Reja, Florian Noichl, André Borrmann, and Ioannis Brilakis. Zero-shot learning with vision-language models for estimating building energy efficiency from street view images. In *Computing in Civil Engineering 2025: Resilient, Robotic, and Educational Systems*, pages 254–262. 2025.
- [42] Deepika Raghu, Iro Armeni, and Catherine De Wolf. Urban-scale facade material mapping from street view images using vision–language models for circular construction planning. *Scientific Reports*, 2026.
- [43] Siyuan Yao, Siavash Ghorbany, Meghan Forstchen, Alexis Korotas, Matthew Sisk, Ming Hu, and Chaoli Wang. Leveraging multimodal llms for building condition assessment from street-view imagery. In *International Symposium on Visual Computing*, pages 219–231. Springer, 2025.
- [44] Yongqing Jiang, Jianze Wang, Xinyi Shen, and Kaoshan Dai. Large language model for post-earthquake structural damage assessment of buildings. *Computer-Aided Civil and Infrastructure Engineering*, 40(31):6324–6342, 2025.
- [45] Xiucheng Liang, Jinheng Xie, Tianhong Zhao, Rudi Stouffs, and Filip Biljecki. Openfacades: an open framework for architectural caption and attribute data enrichment via street view imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 230:918–942, 2025.
- [46] Noelle Law and Yuki Miura. Roofnet: A global multimodal dataset for roof material classification. *arXiv preprint arXiv:2505.19358*, 2025.
- [47] Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9617–9626, 2019.
- [48] Katherine M Collins, Umang Bhatt, and Adrian Weller. Eliciting and learning with soft labels from every annotator. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 10, pages 40–52, 2022.
- [49] Yizhen Yan and Bo Huang. Estimation of building height using a single street view image via deep neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 192:83–98, 2022.
- [50] Asif Azad, Mohammad Sadat Hossain, MD Sadik Hossain Shanto, M Saifur Rahman, and Md Rizwan Parvez. The art of saying "maybe": A conformal lens for uncertainty benchmarking in vlms. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5185–5201, 2026.
- [51] Xiao Liang, Di Wang, Zhicheng Jiao, Ronghan Li, Pengfei Yang, Quan Wang, and Tat-Seng Chua. Uncertainty-driven expert control: Enhancing the reliability of medical vision-language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21144–21154. IEEE, 2025.

- [52] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, 2023.
- [53] Liyao Li, Hao Chen, Jiaming Tian, Wentao Ye, Lirong Gao, Chao Ye, Ningtao Wang, Xing Fu, Yu Cheng, Haobo Wang, et al. Reinforcement learning with verbalized probabilities for llm classification. In *Proceedings of the ACM Web Conference 2026*, pages 7401–7411, 2026.
- [54] Qian Yu, Chaofeng Wang, Barbaros Cetiner, Stella X Yu, Frank McKenna, Ertugrul Taciroglu, and Kincho H Law. Building information modeling and classification by visual learning at a city scale. *arXiv preprint arXiv:1910.06391*, 2019.
- [55] Zongrong Li, Yunlei Su, Hongrong Wang, and Wufan Zhao. Buildingview: Constructing urban building exteriors databases with street view imagery and multimodal large language model. In *International Conference on Spatial Data and Intelligence*, pages 1–19. Springer, 2025.
- [56] EERI. Special Earthquake Report: M6.0 South Napa Earthquake of August 24, 2014. Technical report, Earthquake Engineering Research Institute (EERI), 2014.
- [57] California Seismic Safety Commission. Homeowner’s Guide to Earthquake Safety, 2020.
- [58] Meredith Lochhead, Ádám Zsarnóczy, and Gregory G Deierlein. Exposure matters: A synthesis framework for high-resolution building inventory development. *International Journal of Disaster Risk Reduction*, page 106148, 2026.
- [59] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024.
- [60] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [61] Tobias Groot and Matias Valdenegro-Toro. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 145–171, 2024.
- [62] Xi Wang and Eric Nalisnick. Are vision language models robust to classic uncertainty challenges? *Transactions on Machine Learning Research*, 2026.
- [63] Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision*, pages 148–166. Springer, 2022.
- [64] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Nee-lakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [65] Yang Wu, Shilong Wang, Hao Yang, Tian Zheng, Hongbo Zhang, Yanyan Zhao, and Bing Qin. An early evaluation of gpt-4v (ision). *arXiv preprint arXiv:2310.16534*, 2023.
- [66] Kenneth T Jackson. *Crabgrass frontier: The suburbanization of the United States*. Oxford University Press, 1987.
- [67] Dolores Hayden. *Building suburbia: Green fields and urban growth, 1820-2000*. Vintage, 2004.
- [68] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

- [69] Sebastião Appleton Figueira, Mehrshad Amini, Daniel T Cox, and Andre R Barbosa. Methodology for virtual damage assessment and first-floor elevation estimation: Application to fort myers beach, florida and hurricane ian (2022). *Natural Hazards Review*, 26(2):04025012, 2025.
- [70] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- [71] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibong Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yanzhi Zhu, Mingkun Yang, Zhaozhai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.