Vision-Language Models for Automated Property Feature Extraction in Tax Assessment: A Case Study

Gustavo R. Ribeiro[1], Pedro A. M. Saraiva[1], Luis H. A. Rosa[1], Enzo L. Marques[1], Gustavo L. B. Pereira[1], Pedro M. L. Campos[1] Luiz M. L. Pascoal[1], Sávio S. T. de Oliveira[1]

[1] Instituto de Informática – Universidade Federal de Goiás (UFG)

{ribeirogustavo, pedro.saraiva, luis_alves}@discente.ufg.br

{enzolemes, gustavobueno, campos23}@discente.ufg.br

luizmlpascoal@gmail.com

savioteles@ufg.br

Abstract. *The calculation of property taxes, such as the Urban Building and Land Tax (IPTU) in Brazil, is a critical function for municipal revenue. This process traditionally relies on manual, on-site inspections to assess property characteristics, a task that is costly, time-consuming, and prone to subjectivity. This paper explores the potential of automating this process through the application of state-of-the-art Vision-Language Models (VLMs). We present a novel benchmark to evaluate the capabilities of twelve different VLMs in identifying and classifying specific building features as defined by the municipal legislation of Goiânia, Brazil. Using a dataset of images from public real estate listings and a zero-shot prompting strategy, we tasked the models with extracting 11 distinct construction categories, such as flooring, structure, and finishes. Our results indicate that proprietary models, particularly Google's Gemini 1.5 Pro and Gemini 1.5 Flash, achieve the highest performance, with macro F1-scores of 0.77 and 0.76, respectively. We provide a detailed analysis of model performance across different categories, revealing that while some features like 'Structure' and 'Electrical Installation' are identified with high accuracy, others like 'Sanitary Installation' and 'External Finishes' remain challenging due to their visual subtlety or absence in typical photographs. Our findings demonstrate the significant potential of VLMs to streamline public administration tasks, while also highlighting current limitations and avenues for future research.*

1. Introduction

Property tax represents a fundamental source of revenue for municipalities worldwide, enabling the funding of essential public services such as infrastructure, healthcare, and education. In Brazil, the Urban Building and Land Tax (IPTU) is calculated based on the market value (*valor venal*) of a property. This value is determined by a complex set of characteristics defined in the municipal tax code, including its location, size, and specific construction standards, as stipulated, for instance, in the tax code of Goiânia. Traditionally, the assessment of these characteristics has been a labor-intensive process, requiring on-site inspections by municipal agents. This manual approach is not only

expensive and slow but is also susceptible to human error and subjectivity, leading to potential inconsistencies in tax assessment and disputes from taxpayers.

The recent and rapid advancements in Artificial Intelligence (AI), particularly in the domain of Vision-Language Models (VLMs), offer a transformative opportunity to modernize and automate such administrative processes [Zhang et al. 2024, Radford et al. 2021]. VLMs integrate deep learning models for computer vision and natural language processing, enabling them to comprehend, reason about, and generate text based on visual content. By leveraging publicly available data, such as images from online real estate portals, it may be possible to create a scalable, cost-effective, and consistent system for property characteristic assessment. This could be deployed within various workflows, such as a system to support municipal agents or a service where citizens can request a tax revision, as envisioned by our partner municipality.

This paper addresses the problem of automating the identification of building features for IPTU calculation, framed as a visual classification task governed by legal definitions. Our primary objective is to conduct a comprehensive evaluation of state-of-the-art VLMs on their ability to recognize and classify a predefined set of construction attributes based on the official guidelines stipulated in the Municipal Tax Code of Goiânia, Brazil.

The main contributions of this work are threefold. First, we establish the first benchmark, to our knowledge, for VLM-based property assessment grounded in official legal statutes, presenting a specific application of VLMs in public administration based on a particular dataset. Second, we provide a rigorous comparative analysis of twelve prominent VLMs from the Gemini, Gemma, LLaVA, and Qwen families, evaluating their performance, efficiency, and the impact of hyperparameters like temperature. Third, we offer detailed, category-specific insights into the strengths and weaknesses of current models, highlighting which property features are easily identifiable and which remain challenging, thereby paving the way for future research and practical implementation.

This paper is organized as follows. Section 2 provides the theoretical background on VLMs and discusses related work in AI for real estate. Section 3 details our research methodology, including dataset creation and experimental setup. Section 4 presents and interprets the experimental results in depth. Finally, Section 5 concludes the paper, summarizing our findings, discussing limitations, and proposing directions for future work.

2. Background and Related Work

This section reviews the foundational concepts of Vision-Language Models (VLMs) and surveys related studies that apply artificial intelligence to real estate and property assessment tasks, contextualizing the novelty of our research.

2.1. Vision-Language Models (VLMs)

VLMs represent a paradigm shift in multimodal AI, designed to process and understand information from both images and text simultaneously. They build upon the immense success of Large Language Models (LLMs), which are based on the Transformer architecture and have demonstrated remarkable capabilities in text understanding and generation [Vaswani et al. 2017, Brown et al. 2020]. The core innovation of VLMs is the effective integration of a powerful vision encoder, which has also evolved from Convolutional Neural Networks (CNNs) [He et al. 2016, Jaderberg et al. 2016] to the now-dominant Vision

Transformer (ViT) [Dosovitskiy et al. 2021]. A comprehensive overview of the field can be found in recent surveys [Zhang et al. 2024].

The ViT architecture processes an input image by dividing it into a grid of non-overlapping patches and converting them into a sequence of linear embeddings, analogous to tokenizing text [Mishra et al. 2012]. These visual embeddings are then projected into a shared latent space with text embeddings, allowing a central LLM to reason about them jointly. Foundational models like CLIP (Contrastive Language-Image Pre-training) demonstrated the power of this approach by learning rich visual representations from natural language supervision at a massive scale [Radford et al. 2021]. Subsequent research has explored different architectural strategies. Some models, like LLaVA (Large Language and Vision Assistant) [Liu et al. 2023], effectively align pre-trained, off-the-shelf vision and language models. In contrast, other models, such as Google's Gemini family, are built from the ground up to be natively multimodal, potentially leading to a more seamless and powerful integration of modalities from the earliest stages of training [Team et al. 2023].

2.2. Related Work

The application of machine learning in the real estate sector has seen significant growth, evolving from traditional statistical methods to sophisticated deep learning techniques [Choy and Ho 2023]. Early research focused on price prediction using hedonic models with structured, tabular data [Kok et al. 2017]. More recently, computer vision has been used to automatically extract features from property images to improve valuation accuracy. For instance, Poursaeed et al. [Poursaeed et al. 2018] demonstrated that visual features could be used to directly estimate real estate prices, while Law et al. [Law et al. 2019] showed that combining visual features from CNNs with textual descriptions significantly enhanced the performance of appraisal models.

Specifically within the Brazilian context, Afonso et al. [Afonso et al. 2019] developed an ensemble model using Random Forest and Recurrent Neural Networks to predict housing prices from a large dataset of advertisements. Their work highlights the value of combining multiple data modalities but focuses on price regression rather than granular feature classification. With the advent of VLMs, research has shifted towards a more nuanced, semantic understanding of properties. General-purpose VLM tasks like automatic image captioning [Scoparo and Serapião 2020] have shown the ability of models to generate coherent textual descriptions from visual input. In the real estate domain, Schiappa et al. [Koch et al. 2019] showed that VLMs could be fine-tuned for tasks like visual question answering about specific property attributes. In a related domain, automated construction monitoring has utilized computer vision to track progress and identify materials on-site, a task analogous to our goal of identifying construction features in finished buildings.

However, to the best of our knowledge, the direct application of modern, general-purpose VLMs for automating a legally-defined, granular property assessment for tax purposes is a novel area of research. While previous studies focused on general valuation or aesthetic quality, our work is uniquely constrained by the specific categories and discrete options defined in municipal legislation. This study, therefore, aims to bridge the gap between cutting-edge VLM research and its practical application in public sector governance, evaluating their zero-shot reasoning capabilities on a highly specialized and structured task [Zhang et al. 2024].

3. Methodology

This section outlines the systematic approach employed in this study. We detail the procedures for data acquisition and annotation, the experimental design including the computational environment and model execution protocols, and the metrics used for a rigorous performance evaluation.

3.1. Dataset Curation and Task Definition

The foundation of our benchmark is a custom-curated dataset designed to reflect the diversity of urban housing in Brazil. We sourced images from major Brazilian real estate web portals. A total of 50 distinct property listings were manually selected to form the basis of our dataset. To ensure a representative and varied sample, these listings were carefully chosen to include a balanced mix of standalone houses, apartments in vertical buildings, and homes within horizontal condominiums (gated communities). This diversity is crucial for testing the models' generalization capabilities across different architectural styles and property types.

Each of the 50 listings contained a gallery of photographs, with an average of approximately 20 images per listing. This resulted in a total corpus of approximately 878 images. The images provided comprehensive visual information, covering various perspectives of each property, including exterior facades, living rooms, kitchens, bedrooms, bathrooms, and common areas where applicable. The distribution of classes within the dataset is varied, reflecting real-world occurrences where common features appear frequently, while others are rare. A detailed breakdown of each class is available in the project's public repository.

The core task for the VLMs is a multi-label classification problem, where the labels are defined by the legal framework of the city of Goiânia. Specifically, the categories and their possible values are derived directly from Complementary Law nº 344/2021 [1]. To establish a reliable ground truth, a rigorous annotation process was conducted by two human annotators with expertise in civil construction and real estate appraisal. They independently classified each of the 50 properties across all 11 legally-defined categories. Any discrepancies in their initial assessments were resolved through a consensus discussion to ensure the final labels were accurate and consistent.

3.2. Experimental Design and Execution

Our experimental framework was designed for systematic evaluation and reproducibility. We evaluated a comprehensive set of twelve VLMs from four prominent families: Google's proprietary Gemini series [Saeidnia 2023] ('gemini-2.0-flash', 'gemini-2.5-flash', 'gemini-2.5-pro'); Google's open-source Gemma family [Team et al. 2025] ('gemma3:4b', 'gemma3:12b', 'gemma3:27b'); the LLaVA family [Liu et al. 2023] ('llava:7b', 'llava:13b'); and the Qwen family [Bai et al. 2025] ('qwen2.5:7b', 'qwen2.5:14b', 'qwen2.5:32b').

The selection was based on representing a diverse range of architectures (e.g., natively multimodal vs. connector-based), model sizes, and accessibility (proprietary APIs

[1] https://www.goiania.go.gov.br/html/gabinete_civil/sileg/dados/legis/2021/lc_20210930_000000344.html

vs. open-source), which is a common practice in comparative AI benchmark studies. Proprietary models, such as those in the Gemini family, were accessed via their official Application Programming Interfaces (APIs) using authenticated developer keys. Open-source models were run on our local infrastructure. Our decision to focus on instruction-tuned VLMs over foundational models like CLIP was driven by the task's requirements. The task demanded not just classification but also adherence to a complex, legally-defined structure and output in a specific JSON format, capabilities for which modern VLMs are explicitly designed.

A zero-shot prompting strategy was employed for all experiments. For each property, the full set of its images was provided to the model, along with a structured textual prompt. The prompt instructed the model to act as an expert real estate appraiser and, for each of the 11 construction categories, to select the single most appropriate classification from the legally-defined list of options. To ensure structured and machine-readable outputs, the prompt explicitly required the model to respond in JSON format.[2]

All experiments were executed on a dedicated server equipped with four NVIDIA GeForce RTX 4090 GPUs. The experimental pipeline was implemented in Python 3. The use of multiple GPUs enabled the parallel execution of experiments across different models and temperature settings, significantly reducing the total time required for data collection. For each experimental run (a unique combination of a model and a temperature setting), we systematically processed all 50 properties. The model's predictions, structured in the requested JSON format, were logged for each property. The total execution time for processing the entire dataset was also precisely recorded for each run. Upon completion of the raw data collection, the resulting performance metrics were computed and, along with the raw predictions and logs, stored in structured JSON files for subsequent analysis.

3.3. Evaluation Metrics

The performance of the models was quantitatively evaluated using four standard classification metrics, which are well-suited for this multi-label classification task: accuracy, precision, recall, and F1-Score.

Since our task involves 11 distinct and equally important categories, we calculated these metrics for each category individually. To obtain a single, comprehensive measure of overall model performance, we then computed the macro-average of each metric. Macro-averaging calculates the metric independently for each class and then takes the average, thereby treating all classes equally regardless of their frequency in the dataset. This approach is appropriate for our problem, as accurately identifying a rare but high-value feature is just as important as identifying a common one.

4. Results and Discussion

In this section, we present a comprehensive analysis of our experimental results. The discussion is structured into three main parts. First, we evaluate the overall quality of all twelve Vision-Language Models and the impact of the temperature hyperparameter.

[2]For full transparency and reproducibility, the complete prompt structure, experimental code, specific model configurations, and datasets used in this study are publicly available in `https://anonymous.4open.science/r/Artigo-VLMs-Benchmark-645F`

Second, we analyze the critical trade-off between model accuracy and computational efficiency, a key factor for real-world applicability. Finally, we conduct a granular, category-by-category performance analysis. This three-part evaluation allows for a holistic understanding, covering overall performance, practical deployment considerations, and specific model capabilities. These values were chosen to assess model performance under varying levels of output randomness: 0.0 for deterministic, factual responses; 0.5 for a balance of creativity and predictability; and 1.0 for highly creative outputs.

4.1. Overall Model Quality and Temperature Impact

Our initial analysis focuses on establishing a performance baseline across the twelve evaluated VLMs. Table 1 presents the macro-averaged performance metrics (Accuracy, F1-Score, Precision, and Recall) for each model, tested across three different temperature settings (0.0, 0.5, and 1.0).

The results demonstrate a performance hierarchy among the different model families. The proprietary Gemini models from Google substantially outperform all open-source alternatives. Gemini 2.5 Pro stands out as the most capable model, achieving a peak F1-score of 0.77. It is closely followed by its more agile counterpart, Gemini 2.5 Flash, with a nearly identical F1-score of 0.76. This indicates that the most advanced, commercially available models possess a strong inherent capability for this specialized, legally grounded visual classification task even in a zero-shot setting. This superior performance can likely be attributed to their massive training datasets, more sophisticated and natively multimodal architectures, and extensive fine-tuning for following complex instructions.

A significant performance gap separates these top-tier models from the open-source offerings. The best open-source models, ‘gemma3:27b‘ and ‘gemma3:12b‘, achieve respectable but lower F1-scores of 0.64 and 0.63, respectively. The other open-source families, Qwen and LLaVA, lag further behind, with the popular ‘llava:13b‘ model surprisingly scoring at the bottom of our benchmark with a maximum F1-score of only 0.36.

Regarding the temperature hyperparameter, our findings are consistent across almost all models: lower values (0.0 or 0.5) consistently yield better or equivalent performance. A temperature of 1.0, which encourages creativity and diversity in responses, often degrades performance. This is logical for a factual extraction task like ours. The goal is to map visual evidence to a fixed set of legal categories, a process that requires precision and determinism, not creativity. A higher temperature could cause the model to ”hallucinate” a feature that is not present or to deviate from the provided list of options, thereby resulting in an incorrect classification [Ji et al. 2023].

4.2. Computational Cost

While accuracy is paramount, the computational efficiency (i.e., inference speed) is a critical factor for the practical deployment of any AI system in a public administration setting. Table 2 details the total time each model took to process our entire dataset of 50 properties, serving as a proxy for its response time.

The results reveal a stark and predictable trade-off: higher accuracy comes at the cost of longer processing times. The top-performing model, ‘gemini-2.5-pro‘, is also by

Table 1. Overall Performance Metrics of VLM Models by Temperature. The highest F1-Score for each model is highlighted in bold.

Model	Temp.	Accuracy	F1-Score	Precision	Recall
gemini-2.5-pro	0.00	0.73	**0.77**	0.87	0.73
	0.50	0.72	**0.77**	0.90	0.72
	1.00	0.72	0.76	0.86	0.72
gemini-2.5-flash	0.00	0.72	**0.76**	0.89	0.72
	0.50	0.71	**0.76**	0.89	0.71
	1.00	0.72	**0.76**	0.89	0.72
gemini-2.0-flash	0.00	0.66	**0.70**	0.80	0.66
	0.50	0.65	0.69	0.80	0.65
	1.00	0.64	0.67	0.87	0.64
gemini-2.0-flash-lite	0.00	0.63	**0.66**	0.79	0.63
	0.50	0.62	0.65	0.79	0.62
	1.00	0.60	0.64	0.78	0.60
gemma3:27b	0.00	0.63	**0.64**	0.85	0.63
	0.50	0.62	**0.64**	0.85	0.62
	1.00	0.62	0.63	0.85	0.62
gemma3:12b	0.00	0.61	**0.63**	0.75	0.61
	0.50	0.61	**0.63**	0.75	0.61
	1.00	0.60	**0.63**	0.75	0.60
qwen2.5:32b	0.00	0.55	**0.54**	0.58	0.55
	0.50	0.47	0.51	0.74	0.47
	1.00	0.37	0.45	0.73	0.37
qwen2.5:14b	0.00	0.51	0.49	0.55	
	0.50	0.55	0.53	0.67	0.55
	1.00	0.56	**0.55**	0.64	0.56
qwen2.5:7b	0.00	0.55	**0.53**	0.52	0.55
	0.50	0.46	0.50	0.75	0.46
	1.00	0.43	0.49	0.71	0.43
llava:7b	0.00	0.53	**0.50**	0.48	0.53
	0.50	0.52	0.49	0.48	0.52
	1.00	0.44	0.45	0.48	0.44
llava:13b	0.00	0.37	**0.36**	0.50	0.37
	0.50	0.36	**0.36**	0.40	0.36
	1.00	0.29	0.33	0.47	0.29

Table 2. Execution Time by Model and Temperature. The table is sorted from slowest to fastest model.

Model	Execution Time (s)		
	Temp. 0.0	**Temp. 0.5**	**Temp. 1.0**
gemini-2.5-pro	2174.68	2220.29	2319.11
gemini-2.5-flash	1321.98	1287.55	1377.54
qwen2.5:32b	941.36	915.72	892.31
gemma3:27b	773.66	775.61	779.52
gemma3:12b	709.29	523.59	527.96
qwen2.5:14b	542.72	554.28	572.53
llava:13b	402.00	390.39	365.74
gemini-2.0-flash-lite	345.68	327.73	390.74
gemini-2.0-flash	284.75	284.47	324.48
qwen2.5:7b	258.87	273.56	292.71
llava:7b	232.75	240.97	254.75

far the slowest, requiring over 36 minutes to process the 50 properties. This translates to roughly 43 seconds per property. At the other end of the spectrum, the fastest model, ‘llava:7b‘, completed the entire task in under 5 minutes (about 6 seconds per property) but delivered the second-worst performance.

This analysis is vital for implementation considerations. In a real-world scenario, the choice of model is not just about picking the one with the highest F1-score. If the automated assessment is designed to run overnight on a large batch of tax revision requests, the long inference time of ‘gemini-2.5-pro‘ might be an acceptable trade-off for its superior accuracy. If the VLM is intended to be used as an interactive tool by a human tax assessor to get a ”second opinion” in real-time, a 43-second wait per property is likely impractical. In this context, ‘gemini-2.5-flash‘ offers a compelling balance, providing near-identical accuracy to the ‘pro‘ model but being almost 40% faster. Even faster models like ‘gemma3:27b‘ could be considered if near-instantaneous response is prioritized over maximal accuracy.

Regarding the impact of the `temperature` hyperparameter on efficiency, our analysis reveals a minor and largely inconsistent effect. As observed in Table 2, variations in temperature (0.00, 0.50, 1.00) lead to relatively small fluctuations in execution time for most models. For instance, while `gemini-2.5-pro` shows a slight increase in time at higher temperatures (from 2174.68s to 2319.11s), models like `qwen2.5:32b` actually exhibit a marginal decrease (from 941.36s to 892.31s). Several other models, such as `gemma3:27b` and `gemini-2.0-flash`, maintain remarkably stable execution times across different temperature settings. This contrasts with the more consistent impact of temperature on model accuracy, where lower temperatures (0.00 or 0.50) generally yield superior or equivalent performance for this factual classification task.

4.3. Granular Analysis of Category-Specific Performance

To understand the practical reliability of these models, we must look beyond macro averages and analyze their performance on each of the 11 construction categories. Table

3 provides this detailed breakdown for the top-performing models from the four main families, using their best-performing temperature setting.

Table 3. Comparative Performance of Top VLM Models by Category (Best Temperature Setting for each).

Model	Category	Acc.	F1	Prec.	Recall	Model	Category	Acc.	F1	Prec.	Recall
	External Finishing	0.60	0.72	0.92	0.60		External Finishing	0.04	0.08	0.94	0.04
	Internal Finishing	0.78	0.80	0.83	0.78		Internal Finishing	0.72	0.75	0.79	0.72
	Roofing	0.90	0.95	1.00	0.90		Roofing	1.00	**1.00**	1.00	1.00
	Frames	0.78	0.81	0.84	0.78		Frames	0.84	**0.85**	0.86	0.84
	Structure	0.98	**0.99**	1.00	0.98		Structure	1.00	**1.00**	1.00	1.00
gemini-2.5-pro	Ceiling	0.84	0.87	0.93	0.84	**gemma3:27b**	Ceiling	0.94	**0.91**	0.88	0.94
	Electrical Installation	1.00	**1.00**	1.00	1.00		Electrical Installation	1.00	**1.00**	1.00	1.00
	Sanitary Installation	0.72	**0.84**	1.00	0.72		Sanitary Installation	0.28	0.44	1.00	0.28
	Flooring	0.76	0.74	0.83	0.76		Flooring	0.44	0.35	0.30	0.44
	External Cladding	0.26	0.37	0.74	0.26		External Cladding	0.12	0.06	0.85	0.12
	Internal Cladding	0.30	0.39	0.76	0.30		Internal Cladding	0.52	**0.61**	0.74	0.52
	External Finishing	0.62	**0.73**	0.88	0.62		External Finishing	0.06	0.11	0.94	0.06
	Internal Finishing	0.84	**0.84**	0.84	0.84		Internal Finishing	0.70	0.76	0.83	0.70
	Roofing	0.96	0.98	1.00	0.96		Roofing	0.80	0.89	1.00	0.80
	Frames	0.84	0.82	0.80	0.84		Frames	0.74	0.77	0.79	0.74
	Structure	0.98	**0.99**	1.00	0.98		Structure	0.80	0.89	1.00	0.80
gemini-2.5-flash	Ceiling	0.92	0.90	0.88	0.92	**qwen2.5:32b**	Ceiling	0.74	0.80	0.87	0.74
	Electrical Installation	1.00	**1.00**	1.00	1.00		Electrical Installation	0.28	0.44	1.00	0.28
	Sanitary Installation	0.52	0.68	1.00	0.52		Sanitary Installation	0.00	0.00	0.00	0.00
	Flooring	0.78	**0.76**	0.84	0.78		Flooring	0.36	0.25	0.19	0.36
	External Cladding	0.18	0.24	0.86	0.18		External Cladding	0.68	**0.70**	0.71	0.68
	Internal Cladding	0.28	0.39	0.72	0.28		Internal Cladding	0.02	0.04	0.82	0.02

The results can be categorized into three distinct performance tiers, reflecting the varying challenges posed by different construction elements. The models demonstrate exceptional consistency and reliability in identifying features that are visually prominent, structurally integral, and typically well-represented within their extensive training datasets. Categories such as Structure, Electrical Installation, and Roofing consistently achieve near-perfect F1-scores across the top models. For instance, gemini-2.5-pro and gemini-2.5-flash both reach an F1-score of 1.00 for Electrical Installation and near-perfect scores for Structure (0.99) and Roofing (0.95-0.98). This high performance is attributable to their key identifiers (e.g., concrete pillars, visible power outlets, distinct roof tiles) being common objects clearly discernible in typical real estate photography.

For features like Flooring, Frames, Ceiling, and Internal Finishing, performance is generally good but not flawless (F1-scores typically ranging between 0.75 and 0.90 for Gemini models). The challenges here are rooted in visual ambiguity or the need for nuanced perception. For example, accurately classifying different types of flooring or distinguishing between standard and specialized window frames often requires very fine-grained visual discernment of textures, materials, and subtle design details that can be difficult even for human experts to ascertain from a single 2D photograph.

The models consistently struggle with categories that are either inherently concealed or require extremely fine-grained visual distinction that is not consistently captured in the provided image data. External Cladding and Internal Cladding are prime examples, with F1-scores often falling below 0.40 for most models, indicating significant difficulty in reliably identifying these elements from typical photographs.

The most critical failure case is Sanitary Installation. This category represents a fundamental limitation of pure visual assessment, as plumbing systems are, by design, concealed within walls and floors. The non-zero F1-scores achieved by models like

gemini-2.5-pro (0.84) and gemini-2.5-flash (0.68) in this category warrant closer examination. These scores likely stem from the models inferring the presence of sanitary installations based on visible fixtures (e.g., toilets, sinks, faucets) rather than actual assessment of the concealed plumbing itself. This suggests a form of 'educated guesswork' or indirect inference. In contrast, several open-source models, such as qwen2.5:32b, recorded an F1-score of 0.00 for this category. This indicates a more conservative approach where they explicitly 'admit' their inability to visually verify a concealed feature, rather than attempting to infer its presence. This behavioral difference highlights varied approaches to handling ambiguous or unobservable inputs, with open-source models demonstrating a higher degree of caution when faced with unobservable features. Ultimately, this segment underscores a crucial limitation: VLMs can only reliably assess what is tangibly visible in the provided image data, and relying on inference for concealed elements can lead to misleading performance metrics.

5. Conclusion and Future Work

In this paper, we demonstrated that state-of-the-art Vision-Language Models (VLMs) can automate property feature extraction for tax assessment with a respectable level of accuracy. Our comprehensive benchmark highlighted a clear performance hierarchy, with advanced proprietary models like Gemini 2.5 Pro achieving the highest F1-score of approximately 0.77, significantly outperforming open-source alternatives. Notably, among the open-source models, gemma3:27b emerged as the strongest performer, demonstrating commendable capabilities within its class (F1-score ≈ 0.64), particularly in their ability to follow complex, legally defined classification instructions in a zero-shot setting. We also identified a consistent trade-off between model accuracy and computational efficiency, a critical consideration for practical deployment in public administration. Furthermore, our granular analysis revealed that while VLMs excel at identifying visually prominent features like 'Structure' and 'Electrical Installation', they consistently struggle with concealed elements such as 'Sanitary Installation' or visually subtle features like 'External Cladding'.

Our study provides a benchmark for applying VLMs to a real-world public administration challenge, underscoring their significant potential to streamline traditional, manual processes. However, it also highlights inherent limitations. The current dataset, curated from public real estate listings, may not capture all the nuances required for precise tax assessment, and the zero-shot approach, while demonstrating strong baseline capabilities, can be further improved.

A primary limitation of this study is the potential for data contamination. Given that the VLMs were trained on vast internet corpora, our dataset of public real estate listings may have been included in their training data, potentially leading to inflated performance metrics. Future validation efforts should therefore prioritize the use of private or novel datasets to ensure a more robust assessment of the models' generalization capabilities.

To advance this research, we propose three future works. First, domain-specific fine-tuning of VLMs using a larger, more diverse dataset, potentially including examples from actual municipal tax inspections, could significantly boost performance, especially for challenging categories. Second, the integration of multimodal data beyond images

is crucial. Combining visual information with textual descriptions (e.g., from property advertisements), building blueprints, or even construction permits could provide the necessary context to accurately infer concealed features or ambiguities that are invisible in photographs. Third, exploring human-in-the-loop systems would be a critical next step for practical deployment, leveraging VLMs for initial assessment while flagging ambiguous cases for human review.

6. Acknowledgements

This work was supported by Prefeitura de Goiânia and the National Institute of Science and Technology (INCT) in Responsible Artificial Intelligence for Computational Linguistics and Information Treatment and Dissemination (TILD-IAR) grant number 408490/2024-1.

References

Afonso, B. K. d. A., Melo, L. C., de Oliveira, W. D. G., Sousa, S. B. d. S., and Berton, L. (2019). Housing prices prediction with a deep learning and random forest ensemble. In *Anais do XX Encontro Nacional de Inteligência Artificial e Computacional*, pages 556–567. SBC.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., and Lin, J. (2025). Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. In *Advances in neural information processing systems*, volume 33, pages 1877–1901.

Choy, L. H. and Ho, W. K. (2023). The use of machine learning in real estate research. *Land*, 12(4):740.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Jaderberg, M., Simonyan, K., Vedaldi, A., and Zisserman, A. (2016). Reading text in the wild with convolutional neural networks. *International journal of computer vision*, 116:1–20.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Koch, D., Despotovic, M., Leiber, S., Sakeena, M., Döller, M., and Zeppelzauer, M. (2019). Real estate image analysis: A literature review. *Journal of Real Estate Literature*, 27(2):269–300.

Kok, N., Koponen, E.-L., and Partanen, A.-P. (2017). Big data in real estate? from manual appraisal to automated valuation. *Journal of Portfolio Management*, 43(6):202–211.

Law, S. T., Köse, I. I., Shen, Y., Zhai, X., and Li, S. (2019). House price estimation from visual and textual features. *arXiv preprint arXiv:1902.04944*.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023). Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Mishra, A., Alahari, K., and Jawahar, C. (2012). Scene text recognition using higher order language priors. pages 1–11.

Poursaeed, O., Matera, T., and Belongie, S. (2018). Vision-based real estate price estimation. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Saeidnia, H. R. (2023). Welcome to the gemini era: Google deepmind and the information industry. *Library Hi Tech News*, (ahead-of-print).

Scoparo, M. N. and Serapião, A. B. (2020). Deep learning for automatic image captioning. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 706–717. SBC.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al. (2025). Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Zhang, J., Huang, J., Jin, S., and Lu, S. (2024). Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.