Multimodal Vision-Language Models for Automated Property Feature Extraction: A Comparative Analysis of Image, Text, and Combined Inputs

Pedro M. L. Campos[1], Gustavo R. Ribeiro[1], Enzo L. Marques[1], Gustavo L. B. Pereira[1], Luiz M. L. Pascoal[1], Fernando M. Federson[1], Sávio S. T. de Oliveira[1]

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

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

luizmlpascoal@gmail.com, {federson, savioteles}@ufg.br

Abstract. *Manual property assessment for taxation is costly, time-consuming, and subjective. This paper investigates Vision-Language Models (VLMs) to automate this task through comprehensive evaluation across image-only, text-only, and combined inputs. We evaluated six models from the Gemini and Gemma families on 200 properties from Goiânia, Brazil, classified across 11 legally-defined construction categories using specialized prompting strategies. Our analysis reveals a counterintuitive finding: text-only inputs achieve the highest accuracy, outperforming image-only and matching combined multimodal approaches. This demonstrates that structured textual descriptions contain exceptionally high signal value for legally-defined tasks.*

1. Introduction

Property tax constitutes a primary source of revenue for municipal governments worldwide. This revenue is essential for funding core public services, such as education, infrastructure, and public safety [Force 2022]. This assessment involves evaluating multiple construction characteristics, such as structural systems, finishing materials, electrical installations, and roofing types, each of which contributes to determining a property's taxable value under municipal regulations [Afonso et al. 2013].

Traditionally, classifying these construction features requires labor-intensive on-site inspections by trained municipal agents who manually document each property's attributes according to legal specifications. This manual approach is costly, time-consuming, and susceptible to inconsistencies, creating significant bottlenecks in tax assessment systems. In Brazil, where the Urban Building and Land Tax (IPTU[1]) is calculated based on legally-defined construction standards specified in municipal codes, these challenges are particularly acute [Nascimento and de Jesus Souza 2025].

Recent advances in Vision-Language Models (VLMs) have demonstrated remarkable capabilities in understanding visual and textual content [Zhang et al. 2024]. However, a fundamental question remains: which input modality provides the most reliable

[1]Imposto Predial e Territorial Urbano in Portuguese.

signal for legally-defined property classification? We hypothesize that structured textual descriptions may contain sufficient information for accurate classification, potentially eliminating the need for costly visual inspection systems. This work systematically tests this hypothesis through comparative analysis across three input modalities using models from the Gemini and Gemma families on 200 properties in Goiânia, Brazil.

This work addresses this question through a systematic comparative analysis of VLM performance across three distinct input modalities: image-only, text-only, and combined multimodal inputs. We evaluated models from the Gemini and Gemma families on 200 properties in Goiânia, Brazil, classified across 11 construction categories defined by municipal law. Our analysis reveals a counterintuitive finding: text-only inputs achieve accuracy comparable to or exceeding multimodal approaches, challenging conventional assumptions about the necessity of visual inspection for property assessment.

Our contributions are threefold: (1) the first benchmark evaluating multimodal VLMs performance for property assessment grounded in legal statutes, with comprehensive analysis across all three input modalities; (2) a novel specialized prompting strategy that adapts general-purpose VLMs to domain-specific classification through type-specific instructions, eliminating fine-tuning requirements; and (3) empirical evidence demonstrating that structured textual descriptions can match or exceed the performance of visual and multimodal analysis for legally-structured classification tasks.

2. Related Work

Vision-Language Models represent a paradigm shift in multimodal AI, designed to jointly process and reason about images and text. Building upon the Transformer architecture [Vaswani et al. 2017], VLMs integrate a vision encoder, typically a Vision Transformer (ViT) [Dosovitskiy et al. 2021] that processes images by dividing them into patches and converting them into embeddings analogous to text tokens. These visual and textual embeddings are projected into a shared latent space, enabling unified reasoning across modalities.

Foundational models such as CLIP demonstrated the effectiveness of learning visual representations through natural language supervision on a scale [Radford et al. 2021]. Modern VLMs employ varying architectural strategies: some use connectors to align pre-trained vision and language models [Liu et al. 2023a], while others, such as Google's Gemini family, are natively multimodal from the ground up [Team et al. 2023]. These advances have enabled zero-shot performance on diverse vision-language tasks without task-specific fine-tuning.

Machine learning applications in real estate have evolved from traditional hedonic pricing models using tabular data [Kok et al. 2017] to sophisticated deep learning approaches. Poursaeed et al. [Poursaeed et al. 2018] demonstrated that CNN-extracted visual features could directly estimate property prices, while Law et al. [Law et al. 2019] showed that combining visual and textual features significantly improved the accuracy of the evaluation. In the Brazilian context, Afonso et al. [Afonso et al. 2019] developed ensemble models for price prediction, highlighting the value of multimodal data but focusing on regression rather than feature classification.

Research has shifted towards prompt engineering, a strategy for adapting large models to specialized domains without costly fine-tuning [Liu et al. 2023b]. Pioneered

by models like GPT-3, these methods encode domain knowledge and task examples directly into the prompt, enabling powerful zero-shot and few-shot performance through in-context learning [Brown et al. 2020]. Advanced techniques, such as Chain-of-Thought prompting, further demonstrate how structuring prompts to elicit reasoning can solve complex tasks [Wei et al. 2022]. However, the application of these powerful techniques to legally-structured classification problems, such as those in tax assessment, remains underexplored.

Although a recent study established the first benchmark for applying modern VLMs to property assessment as defined by municipal tax law, its analysis was limited to an image-only, zero-shot approach [Ribeiro et al. 2025]. This initial work leaves a fundamental gap in understanding which data modalities provide the optimal signal for this legally-defined classification task. The present study addresses this gap by conducting the first systematic comparison of image-only, text-only, and multimodal inputs, with direct implications for designing cost-effective systems in resource-constrained public administration contexts.

3. Methodology

This section details our systematic approach to evaluating VLM performance across different input modalities, covering dataset curation, experimental design, our specialized prompting strategy, and evaluation metrics.

3.1. Dataset Curation and Task Definition

Our benchmark comprises 200 property listings from major Brazilian real estate portals, manually selected to ensure balanced representation of standalone houses, apartments, and horizontal condominium homes. Each listing includes multiple photographs providing comprehensive visual documentation.

The task is formulated as multi-label classification: models must classify 11 construction categories with labels derived from Goiânia's Complementary Law nº 344/2021[2]. Categories include: Structure, Windows, Flooring, Ceiling, Electrical Installation, Bathrooms, Internal Cladding, Internal Finishes, External Cladding, External Finishes and Roofing.

Ground truth labels were established through independent annotation by two experts in civil construction and real estate appraisal, with discrepancies resolved through consensus.

3.2. Experimental Design

We evaluated six models spanning proprietary (Gemini 2.5 Pro, Gemini 2.5 Flash, Gemini 2.5 Flash-Lite) and open-source (Gemma 3-27b, Gemma 3-12b, Gemma 3-4b) alternatives in three input conditions:

- **Image-only**: Model receives only property photographs
- **Text-only**: Model receives only textual property descriptions
- **Combined (Image+Text)**: Model receives both photographs and descriptions simultaneously

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

This design isolates each modality's impact and enables direct comparison of their relative contributions to classification accuracy.

3.3. Specialized Prompting Strategy

A critical methodological contribution is our specialized prompting approach. Rather than using a single generic prompt, we developed five tailored prompts based on property typology: one general prompt and four type-specific prompts for apartments, single-family homes, condominium houses, and commercial properties.

Each specialized prompt incorporates three key elements. First, *type-specific tendencies* provide empirically-derived guidance on common patterns. For example, the apartment prompt specifies that structures are "ALWAYS concrete," while the condominium prompt warns that structures "tend to be CONCRETE rather than masonry in 89% of cases."

Second, *disambiguation rules* offer explicit guidance for visually ambiguous categories, clarifying distinctions such as "smooth standard" versus "rough rustic" finishes for external cladding, and when aluminum frames qualify as "standard" versus "premium" based on observable features.

Third, *legal compliance mapping* ensures outputs conform to the 11 categories and valid labels defined in Complementary Law nº 344/2021, preventing invalid classifications.

This approach transforms general-purpose VLMs into domain-specialized classifiers without fine-tuning, which would require extensive labeled data and computational resources. All prompts required JSON-formatted output with BIC classifications, textual justifications, and confidence scores (0-100) for human-in-the-loop workflows. We used temperature=0.1 for deterministic responses. Complete prompt templates are publicly available in the project repository [3].

3.4. Evaluation Metrics

Model performance was evaluated using Accuracy, Precision, Recall, and F1-Score. Since our task involves 11 categories of equal importance in tax assessment, we calculated metrics for each category individually and computed macro-averages. This treats all classes equally regardless of frequency, appropriate because accurately identifying rare but high-value features (such as premium finishes) is as important as correctly classifying common structural elements—both directly impact taxable value under municipal law.

4. Results and Discussion

This section presents a comprehensive analysis of our experimental findings, revealing unexpected patterns in VLM performance across different input modalities and providing insights into optimal deployment strategies for property assessment automation.

4.1. Overall Performance Across Modalities

Our results reveal a surprising and counterintuitive performance hierarchy based on the input data provided to the models. Table 1 consolidates the top-performing configurations and the comparison of the modality.

[3] https://anonymous.4open.science/r/prompts-vlm-erigo

Table 1. Overall Performance: Top Configurations and Modality Comparison

Rank	Model	Input Type	F1	Acc.	Prec.
Top 5 Model Configurations by F1-Score					
1	gemini-2.5-pro	Text+Image	**0.8121**	0.8156	0.8450
2	gemini-2.5-pro	Text-Only	**0.8091**	0.8218	0.8048
3	gemini-2.5-flash	Text-Only	0.8065	0.8218	0.8078
4	gemini-2.5-pro	Image-Only	0.7810	0.7771	0.8159
5	gemma3-27b	Text-Only	0.6878	0.6980	0.6878
Best Model for Each Input Modality					
-	gemini-2.5-pro	Image-Only	0.7810	0.7771	0.8159
-	gemini-2.5-pro	Text-Only	**0.8091**	0.8218	0.8048
-	gemini-2.5-pro	Text+Image	0.8121	0.8156	0.8450

The analysis reveals a critical and unexpected finding: text-only inputs achieve performance nearly equivalent to, and in some model configurations superior to, combined multimodal approaches. The Gemini 2.5 Pro model with text-only input achieves an F1-score of 0.8091, remarkably close to the same model's performance with combined text and image inputs (F1-score of 0.8121). This minimal performance difference of only 0.003 suggests that for this particular legally-defined classification task, the textual property descriptions contain nearly all the information required for accurate classification, with visual data providing only marginal additional value.

Most surprisingly, when examining the Gemini 2.5 Flash model, the combined multimodal approach (F1-score of 0.6557) actually underperforms compared to text-only inputs (F1-score of 0.8065), representing a substantial performance degradation of approximately 19%. This counterintuitive result suggests that for certain model architectures or optimization strategies, the addition of visual information may actually introduce noise or confusion that degrades the classification accuracy rather than enhancing it.

The data clearly illustrate the relative value of different data sources for this specific task. Text-only inputs substantially outperform image-only inputs (F1-score of 0.8091 versus 0.7810, a gain of +0.0281), while the addition of images to text provides only minimal incremental benefit (F1-score improvement of only +0.003). This finding has profound implications for practical deployment strategies and challenges conventional assumptions about the necessity of visual inspection for property assessment.

4.2. Model Architecture Comparison and Performance Scaling

The performance gap between proprietary and open-source models is substantial across all input modalities, and among open-source models, parameter count has a clear impact on performance. Table 2 presents this comprehensive comparison.

The proprietary models consistently outperform their open-source counterparts in all input modalities. This performance advantage can be attributed to their massive training datasets, more sophisticated architectures, and extensive fine-tuning. Among open-source models, Gemma3-27b emerges as the strongest performer, particularly in text-only scenarios (F1-score of 0.6878).

There is a clear positive correlation between model size and classification performance, with the 27B parameter model achieving an F1-score approximately 40% higher

Table 2. Model Architecture Comparison: Proprietary vs. Open-Source and Impact of Model Size

Category	Model	Input Type	Parameters	F1
Proprietary vs. Open-Source Performance				
Image-Only	gemini-2.5-pro (Prop.)	Image	-	**0.7810**
	gemini-2.5-flash-lite (Open)	Image	-	0.5562
Text-Only	gemini-2.5-pro (Prop.)	Text	-	**0.8091**
	gemma3-27b (Open)	Text	27B	0.6878
Text+Image	gemini-2.5-pro (Prop.)	Text+Image	-	**0.8121**
	gemma3-27b (Open)	Text+Image	27B	0.7113
Impact of Model Size (Gemma Family, Text-Only)				
Small	gemma3-4b	Text	4B	0.4888
Medium	gemma3-12b	Text	12B	0.6584
Large	gemma3-27b	Text	27B	**0.6878**

than the 4B version. This suggests that for organizations considering open-source deployment to avoid ongoing API costs, investing in computational resources to run larger models yields significant accuracy improvements that may justify the infrastructure expense.

While accuracy is paramount to ensure fair and consistent tax assessment, practical implementation requires considering computational costs. Proprietary models accessed via cloud APIs incur per-request charges that scale linearly with usage volume. Conversely, open-source models require significant upfront infrastructure investment in specialized hardware such as GPUs, but offer unlimited inference at marginal cost once deployed. For municipal tax assessment, open-source models like Gemma3-27b, despite achieving lower accuracy, can offer better long-term cost-effectiveness for budget-constrained municipalities when combined with human review workflows for properties flagged with low confidence scores.

4.3. Understanding the Multimodal Performance Anomaly

The unexpected underperformance of the Gemini 2.5 Flash model in the combined multimodal condition compared to text-only inputs warrants careful analysis. Several hypotheses may explain this counterintuitive finding.

First, the model's vision-language alignment may be suboptimal for this specific domain, where technical construction terminology and legal categories require precise understanding that the visual modality may confuse rather than clarify. Second, the attention mechanism in multimodal processing may be allocating insufficient weight to the highly informative textual descriptions, instead focusing on visual features that are ambiguous or misleading for legally-defined categories. Third, there may be fundamental limitations in how real estate photography captures the specific features required for tax assessment, with many critical architectural elements either invisible, poorly photographed, or visually indistinguishable in typical property listing images.

This finding suggests that the theoretical advantage of multimodal learning – where complementary information sources should enhance performance – does not automatically materialize in practice for all task types. For highly structured, legally-defined

classification problems where textual descriptions are standardized and comprehensive, the visual modality may provide limited additional signal while introducing significant noise.

4.4. Category-Specific Performance Analysis

A deeper analysis at the individual feature level reveals which property attributes are best identified by each data modality. Table 3 presents the general difficulty in predicting each construction category and the modality-specific performance of Gemini 2.5 Pro.

Table 3. Category-Specific Performance Analysis

Category	Avg F1	Gemini 2.5 Pro F1			Best
	(All Models)	Image	Text	Text+Img	Modality
Electrical Installation	0.8668	0.9354	**0.9800**	0.9082	Text
Structure	0.7798	0.7253	**0.9017**	0.8075	Text
Internal Finishes	0.7762	0.8159	0.8333	**0.8462**	Text+Img
Roofing	0.6929	**0.9170**	0.8448	0.8960	Image
External Finishes	0.6740	0.7692	**0.8333**	0.7692	Text
Flooring	0.6729	0.8537	0.8607	**0.8712**	Text+Img
Windows & Doors	0.6241	0.7889	0.7821	**0.8000**	Text+Img
Bathrooms	0.4919	0.8822	**0.9540**	0.8954	Text
External Cladding	0.4914	0.7154	**0.7241**	0.6852	Text
Internal Cladding	0.4674	**0.7822**	0.6721	0.7808	Image
Ceiling	0.3715	0.6625	0.6095	**0.7386**	Text+Img

The results reveal distinct patterns across different categories of construction. **Text-dominated categories** such as Structure, Electrical Installation, and Bathrooms show substantial improvements with text-only inputs. For Structure, the F1-score increases from 0.73 with images to 0.90 with text alone, likely because textual property descriptions explicitly state construction type using precise terminology such as 'reinforced concrete' or 'masonry', whereas visual evidence of structural systems can be ambiguous or completely hidden behind finished surfaces.

In contrast, **image-dominated categories** like Roofing and Internal Cladding demonstrate better performance with visual data. Roof types, materials, and architectural features are often not detailed in standardized property descriptions but are clearly visible and distinguishable in exterior photographs. Similarly, internal wall coverings, such as tile work, wallpaper, or exposed brick, are highly visual features that may be omitted from text descriptions or described only in vague terms.

Finally, **multimodal-benefiting categories** such as Ceiling and Flooring achieve their best performance when both modalities are available simultaneously. These results suggest that certain architectural features require the complementary strengths of both visual and textual data to resolve classification ambiguities. For example, a ceiling might be partially visible in photographs but require textual context to distinguish between similar-looking materials like gypsum board and PVC panels.

The persistent difficulty of categories such as Ceiling (average F1=0.37) and Internal Cladding (average F1=0.47) in all models and modalities suggests that these features

may require more specialized data collection approaches, additional contextual information, or hybrid human-AI workflows.

4.5. Impact of Specialized Prompting Strategy

Our use of type-specific prompts represents a significant methodological contribution with important practical implications. By encoding domain knowledge and empirically-observed classification patterns directly into the prompting strategy, we effectively created specialized classifiers from general-purpose models without requiring fine-tuning or additional training.

This approach offers several distinct advantages. First, the specialized prompting strategy does not require labeled training data for model adaptation, allowing immediate deployment without extensive data preparation. Second, prompts can be rapidly iterated and updated as classification requirements evolve – when municipal regulations change, new construction materials emerge, or operational experience reveals systematic classification errors, the prompting strategy can be modified within hours or days. Third, explicit rules encoded in specialized prompts make classification logic transparent and auditable, which is critical for public administration applications where decisions must be explainable and legally defensible.

The performance gains from this prompting approach are substantial and measurable. For example, the condominium-specific prompt explicitly warns that the structure "tends to be CONCRETE" in 89% of observed cases, directly addressing a common and costly source of misclassification where traditional masonry construction is incorrectly assumed. Similarly, the commercial property prompt clarifies that an "exposed concrete ceiling is a valid commercial ceiling type," preventing systematic errors where industrial or modern commercial aesthetics with exposed structural elements are misclassified as incomplete construction lacking proper finishing.

4.6. Practical Implications for Deployment

From a practical deployment point of view, our findings challenge conventional wisdom on multimodal AI systems and have important implications for resource allocation and system design. The strong performance of text-only inputs (F1-score of 0.81) relative to combined multimodal approaches suggests that, for this specific task, investing in high-quality textual property descriptions may yield better returns than comprehensive photographic documentation.

For municipalities with limited initial resources or technical infrastructure, a text-only system leveraging property descriptions from existing databases, tax records, or standardized citizen-submitted applications could provide substantial value without requiring image storage infrastructure, bandwidth for transmitting large image files, or the computational overhead of processing visual data.

However, our category-specific analysis reveals that certain features such as Roofing and Internal Cladding do benefit from visual inspection, suggesting that a strategic, feature-specific approach may be optimal. A practical hybrid deployment strategy could use text-only classification as the default for most categories, reserving image analysis for the subset of features where visual information demonstrably improves accuracy. This

selective multimodal approach would balance accuracy, computational efficiency, and operational simplicity.

The specialized prompting approach we developed can be deployed with any of these modality configurations without modification of the underlying models, making the system highly adaptable to varying data availability scenarios in different neighborhoods, types of property or stages of a gradual municipal digitization effort.

5. Conclusion and Future Work

This paper presents the first comprehensive comparative analysis of Vision-Language Models across image-only, text-only, and combined multimodal inputs for legally-defined property assessment. Our research reveals three main findings. First, text-only inputs achieve the highest practical performance (Gemini 2.5 Pro F1=0.81), substantially outperforming image-only analysis (0.78) with minimal difference from combined approaches (0.81). Second, adding visual information provides negligible improvement (+0.003) for best-performing models while degrading performance in some configurations, challenging assumptions about multimodal superiority. Third, our specialized prompting strategy effectively transforms general-purpose VLMs into domain-specialized classifiers without fine-tuning.

Our study has limitations. Textual descriptions from professional listings are highly structured and may contain explicit classification cues, potentially involving data leakage from model training. Our dataset of 200 properties is geographically limited to Goiânia, Brazil. We lack inter-annotator agreement metrics and comparisons with human assessors or traditional baselines, limiting contextualization of absolute performance.

Future research should: (1) validate findings using private municipal data to address contamination concerns; (2) investigate why combined inputs underperform text-only analysis; (3) conduct cost-benefit analysis of proprietary versus open-source deployment at municipal scale; and (4) design hybrid human-AI workflows with confidence-scored predictions and expert review for ambiguous cases. This research demonstrates AI's potential for improving public administration efficiency while revealing that, for legally-structured tasks with high-quality textual descriptions, unimodal text analysis may be sufficient and preferable to complex multimodal approaches.

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.

Afonso, J. R. R., Araújo, E. A., and Nóbrega, M. A. R. d. (2013). O iptu no brasil: um diagnóstico abrangente.

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.

Dosovitskiy et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Force, I. A. I. T. (2022). A review of the methods, applications, and challenges of adopting artificial intelligence in the property assessment office. *Journal of Property Tax Assessment & Administration*, 19(1):2.

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. (2023a). Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Liu, P. et al. (2023b). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

Nascimento, D. A. M. and de Jesus Souza, W. (2025). Da inadimplência à oportunidade de inovação: O potencial do iptu verde para a reforma tributária e a transformação sustentável das cidades. *REVISTA FOCO*, 18(8):e9382–e9382.

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. et al. (2021). Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Ribeiro, G., Teles, S., Saraiva, P., Henrique, L., and Pascoal, L. M. L. (2025). Vision-language models for automated property feature extraction in tax assessment: A comprehensive benchmark. In *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*. SBC.

Team, G. et al. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models.

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*.