https://doi.org/10.3311/PPci.44138

Periodica Polytechnica Civil Engineering

Evaluating Vision-language Models for Zero-shot Room Area Estimation in Floor Plan Images

Nurlan Aliyev[1*], Árpád Barsi[1]

[1] Department of Photogrammetry and Geoinformatics, Faculty of Civil Engineering, Budapest University of Technology and Economics, Műegyetem rkp. 3., H-1111 Budapest, Hungary
[*] Corresponding author, e-mail: nurlan.aliyev@edu.bme.hu

Received: 29 March 2026, Accepted: 03 June 2026, Published online: 05 August 2026

Abstract

Vision-language models (VLMs) have shown strong performance on multimodal reasoning tasks, yet their ability to perform quantitative analysis of technical drawings remains largely unexplored. This study evaluates the instruction-guided zero-shot performance of three open-source VLMs, on the task of estimating room areas from color-coded residential floor plan images. Of the three models tested, two produced quantitatively evaluable structure outputs. A dataset of 100 floor plans, comprising 472 plan-level room-type evaluation comparisons derived from 894 ground-truth room instances, was used to compare predicted and ground truth areas. Results show moderate predictive correlation ($R^2 \approx 0.71$), but substantial estimation errors (MAPE ≈ 43%), with noticeably larger errors for smaller rooms. A classical pixel counting baseline, using color segmentation, achieved MAPE = 5.68% at the room-instance level, and MAPE = 1.58% when evaluated at the floor plan level, highlighting the limitations of the tested VLMs for precise geometric estimation under this controlled setup. These findings suggest that while the evaluated open-source VLMs can approximate spatial proportions in simplified color-coded layouts, their quantitative accuracy remains significantly below that of deterministic image analysis methods.

Keywords

vision-language models, floor plan analysis, zero-shot learning, spatial estimation, architectural AI

1 Introduction

1.1 Motivation and background

Architectural floor plan analysis enables numerous applications, including real estate valuation, building code compliance verification, energy efficiency assessment, and automated design optimization. Many modern deep-learning approaches for floor plan analysis rely on large labeled datasets containing thousands of manually annotated examples [1, 2]. Recent developments in VLMs like LLaVA [3], MiniCPM-V [4], and GPT-4V [5] show impressive zero-shot performance across various visual understanding tasks. This suggests their potential to analyze technical drawings without needing task-specific training.

VLMs integrate visual perception with Natural Language Processing (NLP), enabling models to interpret images through textual instructions and generate structured outputs based on visual input. Instead of creating fixed classifiers as in traditional computer-vision models, VLMs process both text and images.

1.2 Research gap

Prior work on automated floor plan analysis focuses primarily on qualitative tasks. Object detection methods identify architectural elements such as doors and windows [6, 7], semantic segmentation methods classify room types [1, 2], and generative modeling techniques synthesize new layout designs [8, 9]. However, quantitative spatial reasoning tasks, such as estimating room areas from visual input, have received limited attention in VLM literature. This capability is crucial for practical applications that need measurable, actionable results instead of just semantic labels. Although VLMs have demonstrated potential on spatial reasoning benchmarks focusing on object relationships [10], their effectiveness in performing quantitative estimations in technical drawings remains uncertain. On the other hand, soft computing techniques and Artificial Neural Networks (ANN) have been widely established as robust tools for predicting complex quantitative parameters in civil engineering. For example, predictive AI models have been successfully deployed

Cite this article as: Aliyev, N., Barsi, Á. "Evaluating Vision-language Models for Zero-shot Room Area Estimation in Floor Plan Images", Periodica Polytechnica Civil Engineering, 2026. https://doi.org/10.3311/PPci.44138

to estimate the compressive strength of rubberized concrete [11], regionalize base flow indices in hydrology [12], evaluate rock brittleness and dynamic properties [13], and characterize the depth and length of metal cracks *via* inverse eddy current problems [14].

2 Related work

2.1 Traditional ways of floor plan analysis

Early analysis of floor plans used rule-based techniques and manually designed features to understand layouts [15].

Ahmed et al. [16] created wall detection algorithms employing morphological operations and Hough transforms, reaching 78% accuracy on scanned architectural drawings.

Macé et al. [15] proposed graph-based models where rooms are represented as nodes and adjacency relationships as edges, supporting topological queries and spatial reasoning.

Liu et al. [1] integrated Convolutional Neural Networks (CNNs) with Conditional Random Fields (CRFs) for room segmentation, attaining 85% pixel-level accuracy on a dataset of 10,000 labeled floor plans.

Kalervo et al. [2] introduced graph-based parsing techniques, which represent floor plans with nodes and edges, which allow topological analysis.

Dodge et al. [6] proposed methods for detecting walls, doors, and windows in floor plan images through template matching and structural analysis.

de las Heras et al. [7] developed statistical segmentation techniques that combine low-level image features with high-level semantic constraints.

Zeng et al. [17] introduced attention mechanisms for floor plan element detection, boosting small object recognition by 15% over standard CNN architectures.

However, these supervised learning methods require extensive labeled training data and are mainly focused on classification, not on measuring spatial properties.

2.2 VLMs and spatial reasoning

VLMs originated from large-scale vision-text pre-training, as seen in CLIP [18] and ALIGN [19]. CLIP was trained on 400 million image-text pairs collected from the internet, allowing it to transfer knowledge to new visual ideas *via* natural language descriptions.

LLaVA [3] introduced instruction tuning for VLMs with GPT-4 generated data, leading to strong performance on visual questions. It fine-tunes a vision encoder linked to a LLM. The system connects a pre-trained CLIP vision encoder to the LLaMA language model *via* a learned projection layer, allowing it to process visual inputs and produce coherent textual responses.

MiniCPM-V [4] demonstrated that compressed vision transformers can maintain competitive performance with significantly fewer parameters, only 2.6 billion compared to 7-13 billion in larger models, through efficient architecture design and targeted training strategies.

Qwen-VL [20] and InternVL [21] extended VLM capabilities to handle higher-resolution images and multi-image reasoning, which are crucial for analyzing technical documents. GPT-4V [5] showcased strong performance in diagram understanding and scientific figure interpretation, although its proprietary nature restricts reproducibility. Recent research by Liu et al. [10] on VLM spatial reasoning concentrates on qualitative relationships like "object A is left of object B", instead of exact measurements. Additionally, Thrush et al. [22] evaluated VLMs' understanding of spatial relations using the Winoground benchmark, revealing that even the most advanced models find it challenging to make fine-grained spatial distinctions.

2.3 Zero-shot learning paradigms

Lampert et al. [23] introduced attribute-based zero-shot classification, enabling models to identify new object categories by combining learned visual features.

Xian et al. [24] offered a thorough assessment of zero-shot learning methods, differentiating between traditional zero-shot learning, where no target class examples are provided during training, and generalized zero-shot learning, which includes both seen and unseen classes in the test set.

Brown et al. [25] demonstrated that LLMs can achieve competitive performance on various tasks through zero-shot prompting, where natural language instructions enable task specification without gradient updates. This paradigm shift showed that sufficiently large models develop meta-learning capabilities during pre-training. In computer vision, zero-shot learning traditionally involves recognizing novel object categories using semantic attributes [21, 23].

Radford et al. [18] showed that CLIP's vision-language pre-training enables zero-shot transfer to image classification tasks, often matching or exceeding supervised models trained on specific datasets.

VLM zero-shot evaluation has focused on image captioning [26] and visual question answering [3], showing that models can produce accurate descriptions and answer questions about unseen images.

Li et al. [26] found that BLIP-2 performs well in zero-shot settings by linking frozen image encoders to LLMs *via* lightweight trainable modules.

2.4 Architectural datasets

The ResPlan dataset [27], created by Abouagour and Garyfallidis in 2025, marks a major improvement in floor plan resources. It contains 17,107 residential floor plans with vector geometries represented as Shapely [28] polygons and graph structures that encode room adjacency relationships. The dataset sources floor plans from real-world residential properties, capturing diverse architectural styles, room configurations, and layout complexities common in residential construction. See Fig. 1 for a sample of the dataset [27].

Other notable architectural datasets include CubiCasa5K [2] with 5,000 annotated floor plans focusing on room segmentation, RPLAN [29] containing over 80,000 floor plans in raster format, from residential buildings in Asia, and LIFULL HOME'S [30] with over five million floor plans for layout generation research.

3 Methods

3.1 Dataset and ground truth extraction

To ensure diversity in the total habitable area (60–200 m^2, with an average of 120.4 m^2), the number of rooms (ranging from 3 to 7 per plan), and layout complexity, we randomly sampled 100 floor plans from the ResPlan dataset [27] for evaluation. A sample of 100 floor plans with 472 rooms provides sufficient observations for exploratory statistical analysis.

Fig. 1 A sample from ResPlan dataset [27]

During ground truth extraction, we identified a key methodological issue: the floor plans in the ResPlan dataset have inconsistent coordinate systems. The scale factors for converting units to square meters range from 0.00193 to 0.00661 m^2/unit, a 17-fold difference across the dataset. Using a single uniform scale creates errors of five to twelve times in room area calculations, making precise evaluation impossible. To solve this, we developed a per-plan scale factor extraction method (see Fig. 2), dividing the total documented habitable area stored in the metadata by the sum of all room polygon areas, which are computed from the floor plan coordinates. This yields a plan-specific scale factor to convert units to square meters.

We then multiply each room's polygon area by its corresponding scale factor to obtain the actual room area in square meters. This method aligns with realistic residential room sizes, with living rooms averaging 25.6 ± 25.5 m^2, bedrooms 18.4 ± 6.2 m^2, bathrooms 4.8 ± 1.9 m^2, kitchens 10.7 ± 5.5 m^2, and balconies 5.7 ± 3.7 m^2. The reported room-size statistics are computed at the room-instance level, where each polygonal room occurrence (for example bedroom_0, bedroom_1) is treated as one sample.

Floor plans were generated as color-coded PNG images at 200 dpi, with distinct colors for each room type: light grey for living rooms, green for bedrooms, orange for bathrooms, pale blue for kitchens, and darker gray for balconies. The use of color-coded floor plans was an intentional controlled design choice; by removing the room detection and semantic classification sub-tasks, we isolate the spatial area estimation component, allowing us to evaluate whether VLMs can estimate relative spatial extents of clearly defined regions. This controlled setting provides a clean lower bound on VLM spatial estimation capability before tackling the harder, fully unconstrained setting of raw architectural drawings.

3.2 VLM configuration

We evaluated three open-source VLMs, namely LLaVA-LLaMA 3, MiniCPM-V 2.6 and BakLLaVa *via* Ollama interface. Ollama allows users to download the base models, create specialized versions *via* system prompt, and run them on localhost, without relying on cloud services or proprietary APIs.

All models were accessed by local Ollama inference at a temperature of 0.1 to produce deterministic and reproducible outcomes for comparison. The temperature parameter affects the randomness of the model response. We selected these three models to represent the different architectural

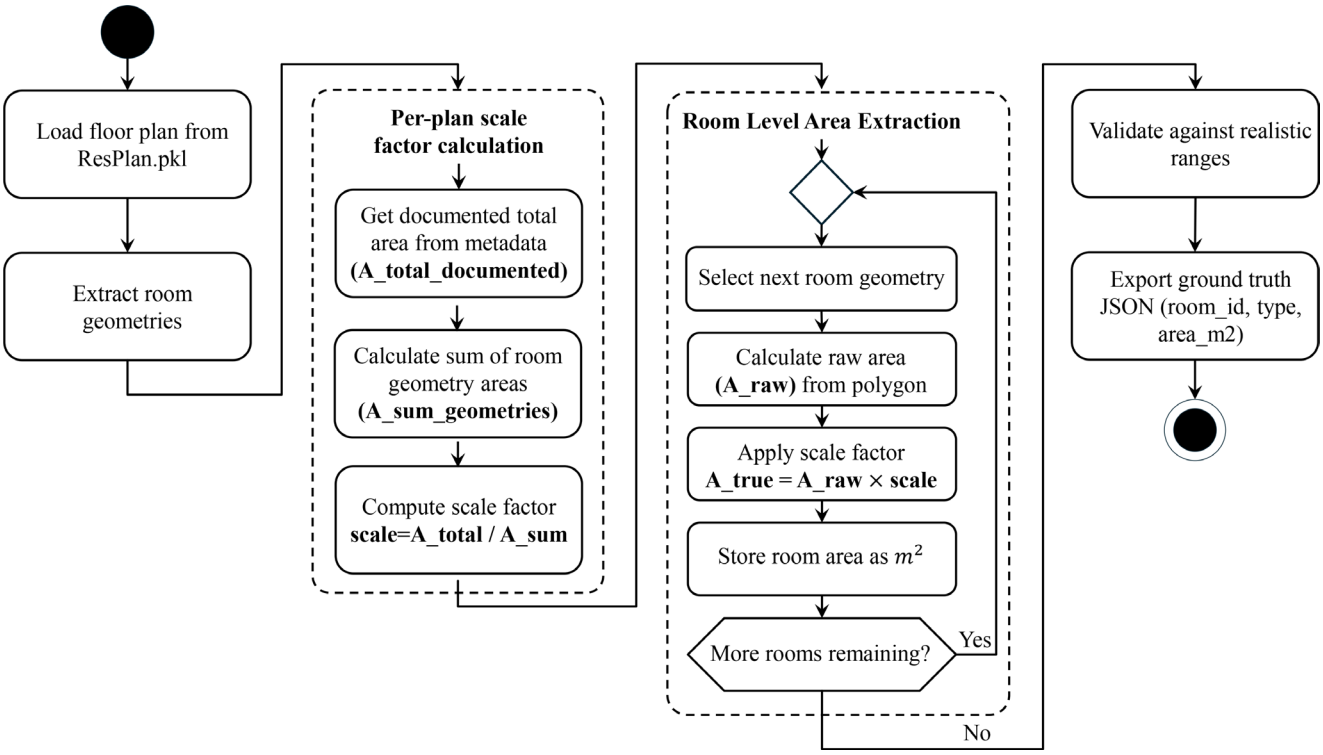


Fig. 2 Per-plan scale factor extraction method

approaches: LLaVA-LLaMA 3 combines the popular CLIP vision encoder with Meta's Llama 3 language model, MiniCPM-V highlights the efficient use of parameters by compression in the architecture, and BakLLaVA provides an alternative LLaVA implementation for reliability testing. Table 1 shows the characteristics of the models and the rationale for their selection.

3.3 Zero-shot prompting strategy

Models received a detailed 158-line (comments and empty lines included) system prompt that specified the task, methodology, domain knowledge, and output format. The prompt contained five key elements that structure the analytical approach of the model. First, we defined the task as the extraction of quantitative spatial information, namely, room areas and interconnection relationships, from floor plans. Secondly, we provided methodological guidance on the visual grid method for estimating area, where models should conceptually divide the floor plan into a grid of 10 × 10 cells with 100 cells occupied to estimate the percentage coverage of the total floor area of each room. Third, we embedded domain knowledge by explicitly mapping colors to room types, eliminating the need for models to perform room-type classification. Fourth, we specified the output format as structured JSON with nodes representing individual rooms (each node including id, type, and estimated_area_percent fields) and edges representing physical connections between rooms, such as doorways or open transitions. The edge detection was an initial idea, yet we focused more on the spatial information extraction rather than graph extraction ability. Fifth, we imposed constraints requiring models to analyze actual image content without producing template placeholder values, ensuring that every output represents genuine analysis of the provided floor plan image and that models must identify all visible rooms present in the image.

The complete system prompt, including the full output schema and visual grid instructions, is provided in the Supplement.

The system prompt provides task instructions and domain knowledge, but contains no labeled training examples, no example floor plans with correct area measurements

Table 1 VLMs evaluated in this study

Model	Architecture	Parameters	Selection rationale
LLaVA-LLaMA 3	CLIP ViT + Llama 3	~8B	Widely adopted, strong instruction-following
MiniCPM-V 2.6	Compressed ViT + LLM	~2.6B	Efficient architecture, competitive performance
BakLLaVA	LLaVa variant	~7B	Alternative LLaVA implementation for comparison

are shown to the model. This configuration constitutes zero-shot learning following established conventions from GPT-3 [25] and LLaVA [3], where natural language task descriptions enable generalization to novel tasks without task-specific training data or gradient updates.

Our inference protocol used a 300-second timeout per image to accommodate complex floor plans requiring extended processing time, allowed a maximum of 2 retry attempts for failed inferences to handle transient errors, and applied strict JSON schema validation to filter malformed outputs and ensure data quality.

3.4 Prediction parsing and room matching

Each model was prompted to return a structured JSON object containing a list of nodes, where each node specifies a room identifier, room type, and estimated area as a percentage of total floor plan area. Outputs were validated against three criteria:

1. the response must be parseable as valid JSON;
2. all node entries must contain the required fields (id, type, estimated_area_percent);
3. no field may contain placeholder values (e.g., literal "%" or "roomtype").

Responses failing validation were retried once; outputs failing on the second attempt were marked as unusable and excluded from quantitative evaluation.

Under these criteria, LLaVA-LLaMA 3 and MiniCPM-V 2.6 produced valid structured outputs for all 100 floor plans. BakLLaVA consistently returned template placeholders rather than genuine predictions across all 100 plans and we excluded it from quantitative evaluation.

For room matching, predictions were aligned to ground truth by room type rather than by individual instance. The ground truth dataset contains 894 room polygon instances across 100 floor plans (165 living rooms, 236 bedrooms, 263 bathrooms, 99 kitchens, 131 balconies). Because VLMs report area as a percentage of total plan area per room type, all ground-truth instances of the same type within a plan were aggregated into a single comparison point by summing their areas. This yields one (predicted, ground-truth) pair per room type per floor plan, resulting in 472 matched comparisons in total (100 plans × up to 5 room types, minus 27 plans containing no balcony and 1 plan containing no kitchen). Plans where a room type was absent from the ground truth were excluded from that room type's evaluation.

3.5 Evaluation metrics

We evaluate model performance using Mean Absolute Percentage Error (MAPE) [31], Root Mean Square Error (RMSE) [32], Mean Absolute Error (MAE), coefficient of determination (R^2) [33] and Pearson Correlation Coefficient (r).

Unlike the R^2 metric used in linear regression, which is constrained to the interval [0,1], the R^2 metric we employ for evaluating predictions (rather than fitted models) can indeed produce negative values when the sum of squared residuals exceeds the total sum of squares [33]. This occurs when model predictions are systematically biased or poorly calibrated, causing them to vary from ground truth in ways that increase variance rather than explaining it.

Two evaluation levels are used: Room-level metrics, computed per individual room, and Plan-level metrics aggregating predictions across each floor plan and computing the error relative to the total plan area.

3.6 Statistical hypothesis testing

To assess systematic bias, we performed a paired t-test comparing predicted and ground truth room areas. Statistical significance was evaluated at $\alpha = 0.05$.

3.7 Implementation details

The evaluation pipeline was implemented in Python 3.10 [34] using Shapely 2.0.1 [28], GeoPandas 0.14.0 [35], SciPy 1.11.0 [36] and Matplotlib 3.7.1 [37]. VLM inference (see Fig. 3) was performed locally using Ollama server version 0.1.32 [38], running on localhost port 11434. Source code and processed data will be released upon publication.

4 Results

4.1 Overall performance comparison

Table 2 reports overall performance aggregated across all 472 plan-level room-type comparisons. The pixel count baseline achieved the best accuracy (MAPE 5.68 ± 7.21%, RMSE 2.09 m^2, MAE 1.01 m^2, R^2 0.990, $r = 0.995$), substantially outperforming both VLMs under this color-coded setup. Among VLMs, MiniCPM-V 2.6 achieved numerically lower RMSE and MAE, while LLaVA-LLaMA 3 achieved numerically lower mean MAPE; no formal model comparison test was conducted given the high within model variability. BakLLaVA consistently returned template placeholder output across all 100 plans and we excluded it from quantitative evaluation (see Section 3.4). See Fig. 4 (a) for a scatter plot visualization of the predicted versus actual relationship.

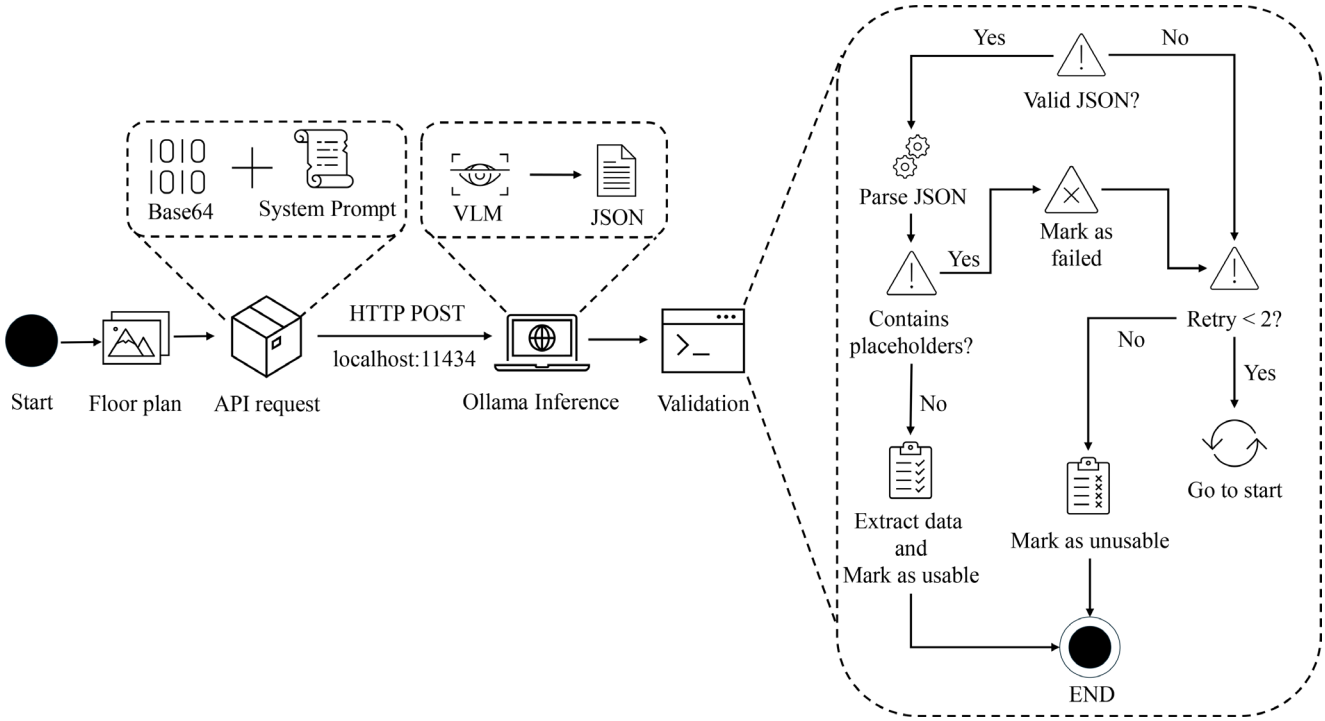


Fig. 3 VLM inference pipeline

Table 2 Overall model performance on 100 residential floor plans (472 room comparisons, plan-level evaluation)

Metric	LLaVA-LLaMA 3	MiniCPM-V 2.6	Pixel count baseline	BakLLaVA
↓ MAPE (%)	41.07 ± 36.91	43.72 ± 51.52	5.68 ± 7.21	N/A
↓ RMSE (m^2)	12.59	11.29	2.09	N/A
↓ MAE (m^2)	8.73	7.91	1.01	N/A
↑ R^2	0.649	0.718	0.990	N/A
↑ Pearson r	0.826*	0.861*	0.995*	N/A

* $p < 0.005$
↓ : Lower is better
↑ : Higher is better

4.2 Room-type-specific performance analysis

Table 3 presents performance breakdown by room type, revealing noticeable variation in VLM accuracy across different architectural spaces. Across all room types, the pixel count baseline yields the lowest MAPE and highest R^2, while VLM performance remains weaker, especially for rooms with small areas.

Large rooms, including living rooms and bedrooms, achieved the strongest accuracy with MAPE values of 21–22%. Strong positive R^2 values ranging from 0.605 to 0.740 indicate that VLMs can reliably estimate areas for spacious architectural elements. Medium-sized rooms such as bathrooms showed moderate performance with 26% MAPE and $R^2 = 0.402$, suggesting partial but imperfect area estimation capability. Small rooms, including kitchens and balconies, exhibited poor performance with MAPE exceeding 56% and, crucially, negative R^2 values. This represents a failure mode where the model's area estimation ability breaks down. See Fig. 4 (b) for error distribution visualization by room type.

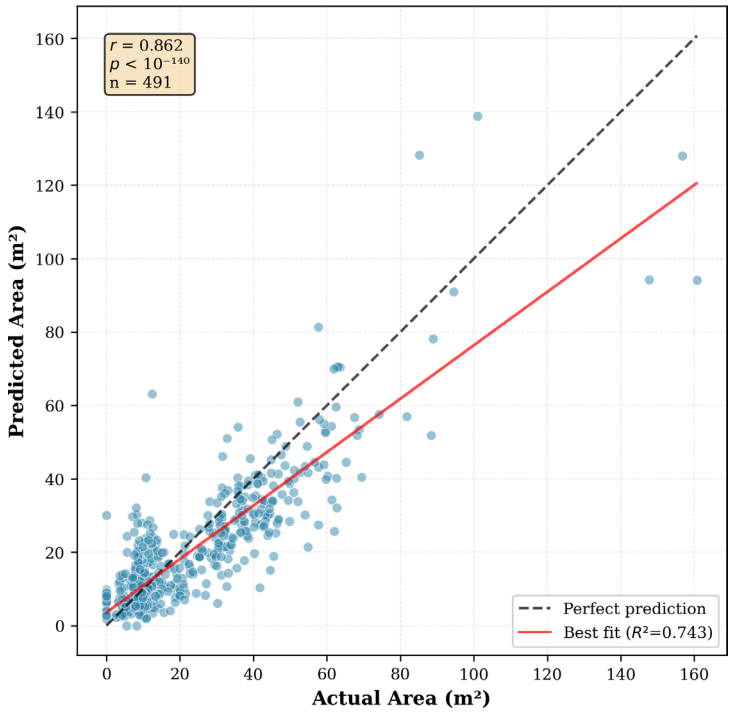
A sharp degradation in performance is observed below approximately 15 m^2, suggesting a size-dependent limitation in the models' estimation capabilities.

4.3 Room size dependency

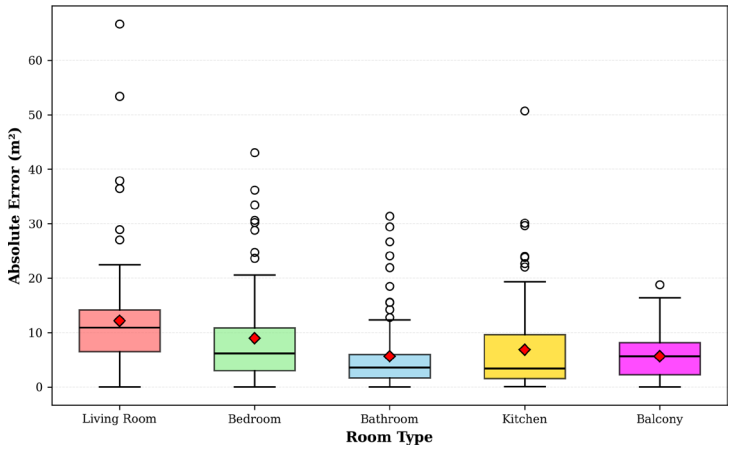
In the relationship between room size and prediction accuracy, we categorized rooms by area, which is represented in Table 4.

Prediction error increases as room size decreases, with a strong inverse correlation between MAPE and room area ($r = -0.89$). See Fig. 5 for a visualization of this relationship. This trend indicates that model accuracy is strongly dependent on the relative spatial extent of the target region.

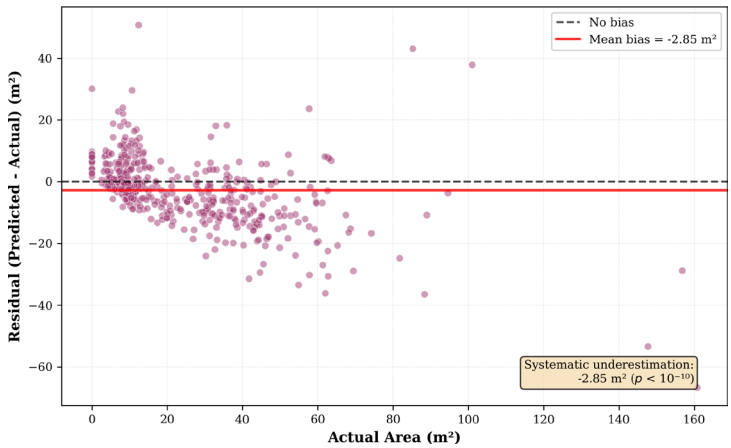
Negative R^2 occurred in 50% of model-room combinations, with MiniCPM-V on kitchens showing failure at $R^2 = -2.41$. When the actual kitchen size increased, the



(a)



(b)



(c)

Fig. 4 Comparison, error and bias plots: (a) Scatter plot of predicted and ground truth room areas for MiniCPM-V; (b) Error boxplots by room type; (c) Residual plot showing systematic bias

Table 3 Room type area estimation performance across models. Mean area is reported at room-instance level (averaged across all ground-truth polygon instances of that type). Error metrics are computed at the plan level, with one aggregated (predicted, ground-truth) comparison per room type per floor plan. N-GT = total ground-truth polygon instances in the dataset; N-eval = number of floor plans included in evaluation for that room type (plans lacking a given room type are excluded). Lower MAPE indicates better accuracy, higher R^2 indicates better fit

Room type	N-eval	N-GT	Mean area (m²)	LLaVA-LLaMA 3 (MAPE, R^2)	MiniCPM-V 2.6 (MAPE, R^2)	Pixel count baseline (MAPE, R^2)
Living room	100	165	25.64 ± 25.61	21.02, 0.740	27.27, 0.477	1.58, 0.995
Bedroom	100	236	18.41 ± 6.18	41.82, −0.273	21.51, 0.605	1.50, 0.996
Bathroom	100	263	4.77 ± 1.92	25.99, 0.402	39.33, −0.057	11.03, 0.771
Kitchen	99	99	10.71 ± 5.52	56.57, −0.571	66.77, −2.409	3.78, 0.991
Balcony	73	131	5.69 ± 3.67	67.14, −0.430	71.43, 0.020	12.27, 0.966

Table 4 Performance by room size category (VLMs *vs*. baseline)

Size category	Area range	Room types	MAPE (%)-VLMs	R^2-VLMs	MAPE (%) - baseline	R^2-baseline
Large	>40 m²	Living, bedroom	21.02–41.82	−0.27–0.74	1.50–1.58	0.995–0.996
Medium	10–15 m²	Bathroom, kitchen	25.99–66.77	−2.41–0.40	3.78–11.03	0.771–0.991
Small	<10 m²	Balcony	67.14–71.43	−0.43–0.02	12.27	0.966

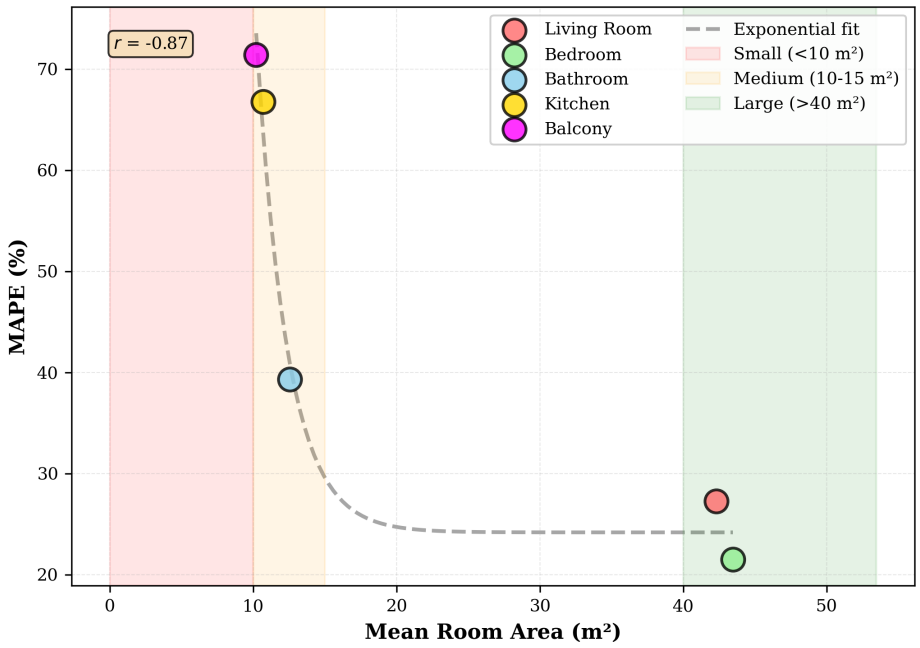


Fig. 5 Room size dependency analysis

predicted size often decreased, and *vice versa*. This counterintuitive behavior suggests fundamental limitations in VLM's fine-grained quantitative area estimation for small architectural elements. Although these explanations remain hypotheses requiring controlled architectural experiments.

4.4 Systematic bias analysis

Table 5 presents results from paired *t*-tests evaluating systematic bias in area estimation.

Both models exhibited statistically significant systematic underestimation of approximately 3 m² with *p*-values far below conventional significance thresholds at $p < 0.001$. The consistency of this bias across both tested architectures (LLaVA-LLaMA 3 and MiniCPM-V) suggests a consistent pattern in spatial estimation behavior rather than a model-specific calibration issue. See Fig. 4 (c) for visualization of this systematic negative bias across all room sizes. This persistent underestimation may stem from

Table 5 Systematic bias test results (paired *t*-test, $n = 472$ room comparisons)

Model	Mean bias (m²)	Standard error	*t*-statistics	Significant ($p < 0.001$)
LLaVA-LLaMA 3	−3.07	0.563	−5.458	Yes
MiniCPM-V	−3.27	0.498	−6.561	Yes
Pixel count baseline	−0.030	0.096	−0.313	No

VLMs' pre-training on natural images, where estimating smaller is conservative (objects are often partially cut off), whereas floor plans show complete room extents.

4.5 Statistical hypothesis testing

Results of paired t-tests for systematic bias presented in Table 5 show highly significant underestimation for both models with $p < 0.001$. We reject the null hypothesis of zero mean error and confirm the presence of systematic bias requiring calibration. The negative t-statistics indicate consistent underestimation rather than random errors centered on zero. The large absolute values of t-statistics (5.458 and 6.561) suggest that this bias is highly robust and would be detected in any similarly sized sample from the same population.

5 Discussion

The results demonstrate that the tested VLMs can approximate room areas from color-coded floor plan representations for large rooms, but with limited accuracy and strong dependence on room size. Performance is not uniform across room types: both models show positive predictive correlation for living rooms and, to a lesser extent, bedrooms, the two largest room categories. For smaller room types, including kitchens, balconies, and bathrooms, at least one model produces negative R^2 values, indicating that predictions for these room types are less informative than a naive mean-area predictor. The aggregate R^2 values reported in Table 2 (0.649 and 0.718) reflect the large-room performance and should not be taken as representative of overall spatial estimation capability.

The comparison with the pixel-based baseline highlights this limitation clearly. The baseline achieves near-perfect accuracy ($R^2 = 0.990$) through direct access to geometric information, whereas the evaluated VLMs rely on indirect visual estimation *via* the conceptual grid method. This gap is not marginal: the baseline's MAPE of 5.68% is approximately seven times lower than the VLMs' median MAPE of ~32%. In the tested zero-shot color-coded setup, the evaluated open-source VLMs do not provide a competitive alternative to deterministic pixel counting for quantitative floor plan area estimation.

A consistent pattern across all experiments is the strong dependence of prediction accuracy on room size. Larger rooms are estimated with relatively lower error, while performance deteriorates significantly for smaller regions. The analysis in Section 4.3 shows a pronounced increase in error below approximately 15 m^2, suggesting that small spatial regions represent a critical limitation for the tested models.

The stronger degradation in MAPE for small rooms may partly reflect the mathematical behavior of percentage-based error metrics rather than a purely perceptual limitation. If the absolute estimation error remains approximately constant across room sizes, which the MAE values in Table 3 suggest is plausible, the same absolute deviation produces a proportionally much larger MAPE for a 5 m^2 balcony than for a 40 m^2 living room. The concurrent reporting of MAE and RMSE alongside MAPE in this study mitigates this interpretive risk, as the absolute error metrics do not exhibit the same size dependency. Future work should examine scale-dependent error more rigorously, for example through absolute-error regression against room area or calibration curve analysis, to separate metric-induced inflation from genuine perceptual degradation.

In addition to size-dependent effects, both evaluated models exhibit a systematic underestimation bias of approximately 3 m^2. This bias appears consistently across room types and floor plans, indicating that it is not associated with specific spatial configurations. Although the origin of this bias cannot be determined from the current experiments, its presence further limits the reliability of the predictions.

It is important to note that the use of color-coded floor plans simplifies the visual interpretation task by removing the need for room detection and semantic segmentation. The results, therefore, isolate the spatial estimation component of the problem. While this controlled setup allows clearer analysis of model behavior, it does not fully represent real-world architectural drawings, where additional sources of error are expected.

Overall, the findings indicate that the tested open-source VLMs, operating in a zero-shot setting on color-coded floor plan images, possess limited capability for quantitative spatial estimation and remain substantially less accurate than the deterministic pixel-counting baseline evaluated here. These conclusions apply to the specific controlled setup described and should not be generalized to all VLM architectures or all floor plan representations without further investigation.

6 Conclusions

This study evaluated the ability of VLMs to estimate room areas from floor plan images in a zero-shot setting. Using a dataset of color-coded residential floor plans, the performance of three open-source VLMs (LLaVA-LLaMA 3, MiniCPM-V 2.6, and BakLLaVA) was assessed; of these, two produced quantitatively evaluable outputs and were compared to a pixel-counting baseline.

The results show that the evaluated VLMs can approximate room areas with moderate correlation to ground truth data; however, their accuracy remains substantially lower than that of the classical baseline. Prediction performance is strongly dependent on room size, with reliable estimates primarily limited to larger regions. A consistent underestimation bias was also observed across models.

The experimental setup, based on simplified color-coded floor plans, isolates the spatial estimation task but does not capture the full complexity of real architectural drawings. Future work should therefore extend the evaluation to raw floor plans, where room boundaries and semantics must be inferred, and investigate methods to improve model performance for small spatial regions.

References

[1] Liu, C., Wu, J., Kohli, P., Furukawa, Y. "Raster-to-Vector: Revisiting Floorplan Transformation", In: 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2214–2222. ISBN 978-1-5386-1033-6 https://doi.org/10.1109/ICCV.2017.241

[2] Kalervo, A., Ylioinas, J., Häikiö, M., Karhu, A., Kannala, J. "CubiCasa5K: A Dataset and an Improved Multi-task Model for Floorplan Image Analysis", In: Felsberg, M., Forssén, P.-E., Sintorn, I.-M., Unger, J. (eds.) Image Analysis, Norrköping, Sweden, 2019, pp. 28–40. ISBN 978-3-030-20204-0 https://doi.org/10.1007/978-3-030-20205-7_3

[3] Liu, H., Li, C., Wu, Q., Lee, Y. J. "Visual Instruction Tuning", In: Advances in Neural Information Processing Systems 36, New Orleans, Louisiana, USA, 2023, pp. 34892–34916. ISBN 9781713899921 [online] Available at: https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf [Accessed: 26 January 2026]

[4] Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., …, Sun, M. "MiniCPM-V: A GPT-4V Level MLLM on Your Phone", [preprint] arXiv, arXiv:2408.01800v1, 03 August 2024. https://doi.org/10.48550/arXiv.2408.01800

[5] OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., …, Zoph, B. "GPT-4 Technical Report", [preprint] arXiv, arXiv:2303.08774v6, 04 March 2024. https://doi.org/10.48550/arXiv.2303.08774

[6] Dodge, S., Xu, J., Stenger, B. "Parsing floor plan images", In: 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA), Nagoya, Japan, 2017, pp. 358–361. ISBN 978-1-5386-0495-3 https://doi.org/10.23919/MVA.2017.7986875

[7] de las Heras, L.-P., Ahmed, S., Liwicki, M., Valveny, E., Sánchez, G. "Statistical segmentation and structural recognition for floor plan interpretation", International Journal on Document Analysis and Recognition (IJDAR), 17(3), pp. 221–237, 2014. https://doi.org/10.1007/s10032-013-0215-2

[8] Shabani, M. A., Hosseini, S., Furukawa, Y. "HouseDiffusion: Vector Floorplan Generation via a Diffusion Model with Discrete and Continuous Denoising", In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 5466–5475. ISBN 979-8-3503-0130-4 https://doi.org/10.1109/CVPR52729.2023.00529

[9] Hu, R., Huang, Z., Tang, Y., Van Kaick, O., Zhang, H., Huang, H. "Graph2Plan: learning floorplan generation from layout graphs", ACM Transactions on Graphics, 39(4), 118, 2020, https://doi.org/10.1145/3386569.3392391

[10] Liu, F., Emerson, G., Collier, N. "Visual Spatial Reasoning", Transactions of the Association for Computational Linguistics, 11, pp. 635–651, 2023. https://doi.org/10.1162/tacl_a_00566

[11] Bachir, R., Sidi Mohammed, A. M., Habib, T. "Using Artificial Neural Networks Approach to Estimate Compressive Strength for Rubberized Concrete", Periodica Polytechnica Civil Engineering, 62(4), pp. 858–865, 2018. https://doi.org/10.3311/PPci.11928

[12] Jolánkai, Z., Koncsos, L. "Base Flow Index Estimation on Gauged and Ungauged Catchments in Hungary Using Digital Filter, Multiple Linear Regression and Artificial Neural Networks", Periodica Polytechnica Civil Engineering, 62(2), pp. 363–372, 2018. https://doi.org/10.3311/PPci.10518

[13] Xie, Y., Wang, L., Gu, Y., Gu, X., Chen, S., Khajehzadeh, M., Hosseini, S. "Prediction of Rock's Brittleness and Dynamic Properties Utilizing Effective Artificial Intelligence Approaches", Periodica Polytechnica Civil Engineering, 68(3), pp. 946–960, 2024. https://doi.org/10.3311/PPci.23156

[14] Benissad, S., Touati, M., Chabaat, M. "Artificial Neural Networks for Inverse Problems in Damage Detection using Computational and Experimental Eddy Current", Periodica Polytechnica Civil Engineering, 67(1), pp. 1–9, 2023. https://doi.org/10.3311/PPci.20550

[15] Macé, S., Locteau, H., Valveny, E., Tabbone, S. "A system to detect rooms in architectural floor plan images", In: Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, Boston, MA, USA, 2010, pp. 167–174. ISBN 978-1-60558-773-8 https://doi.org/10.1145/1815330.1815352

[16] Ahmed, S., Liwicki, M., Weber, M., Dengel, A., "Improved Automatic Analysis of Architectural Floor Plans", In: 2011 International Conference on Document Analysis and Recognition, Beijing, China, 2011, pp. 864–869. ISBN 978-1-4577-1350-7 https://doi.org/10.1109/ICDAR.2011.177

[17] Zeng, Z., Li, X., Yu, Y. K., Fu, C.-W. "Deep Floor Plan Recognition Using a Multi-Task Network with Room-Boundary-Guided Attention", In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 9095–9103. ISBN 978-1-7281-4803-8 https://doi.org/10.1109/ICCV.2019.00919

[18] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, A., …, Sutskever, I. "Learning Transferable Visual Models From Natural Language Supervision", [preprint] arXiv, arXiv:2103.00020, 26 February 2021. https://doi.org/10.48550/arXiv.2103.00020

[19] Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., Duerig, T. "Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision", [preprint] arXiv, arXiv:2102.05918, 11 June 2021. https://doi.org/10.48550/arXiv.2102.05918

[20] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J. "Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond", [preprint] arXiv, arXiv:2308.12966, 13 October 2023. https://doi.org/10.48550/arXiv.2308.12966

[21] Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., …, Dai, J. "Intern VL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks", In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2024, pp. 24185–24198. ISBN 979-8-3503-5300-6 https://doi.org/10.1109/CVPR52733.2024.02283

[22] Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., Ross, C. "Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality", In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 5228–5238. ISBN 978-1-6654-6946-3 https://doi.org/10.1109/CVPR52688.2022.00517

[23] Lampert, C. H., Nickisch, H., Harmeling, S. "Attribute-Based Classification for Zero-Shot Visual Object Categorization", IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(3), pp. 453–465, 2014. https://doi.org/10.1109/TPAMI.2013.140

[24] Xian, Y., Lampert, C. H., Schiele, B., Akata, Z. "Zero-Shot Learning—A Comprehensive Evaluation of the Good, the Bad and the Ugly", IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(9), pp. 2251–2265, 2019. https://doi.org/10.1109/TPAMI.2018.2857768

[25] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., …, Amodei, D. "Language Models are Few-Shot Learners", [preprint] arXiv, arXiv:2005.14165, 22 July 2020. https://doi.org/10.48550/arXiv.2005.14165

[26] Li, J., Li, D., Savarese, S., Hoi, S. "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models", [preprint] arXiv, arXiv:2301.12597, 15 June 2023. https://doi.org/10.48550/arXiv.2301.12597

[27] Abouagour, M., Garyfallidis, E. "ResPlan: A Large-Scale Vector-Graph Dataset of 17,000 Residential Floor Plans", [preprint] arXiv, arXiv:2508.14006, 19 August 2025. https://doi.org/10.48550/arXiv.2508.14006

[28] Sean Gillies, Shapely Contributors "Shapely, (2.0.1)", [computer program] Available at: https://shapely.readthedocs.io/en/2.0.1/index.html [Accessed: 03 November 2025]

[29] Wu, W., Fu, X.-M., Tang, R., Wang, Y., Qi, Y.-H., Liu, L. "Data-driven interior plan generation for residential buildings", ACM Transactions on Graphics, 38(6), 234, 2019. https://doi.org/10.1145/3355089.3356556

[30] Nauata, N., Chang, K.-H., Cheng, C.-Y., Mori, G., Furukawa, Y. "House-GAN: Relational Generative Adversarial Networks for Graph-Constrained House Layout Generation", In: 16th European Conference on Computer Vision, Glasgow, UK, 2020, pp. 162–177. ISBN 978-3-030-58451-1 https://doi.org/10.1007/978-3-030-58452-8_10

[31] de Myttenaere, A., Golden, B., Le Grand, B., Rossi, F. "Mean Absolute Percentage Error for regression models", Neurocomputing, 192, pp. 38–48, 2016. https://doi.org/10.1016/j.neucom.2015.12.114

[32] Chai, T., Draxler, R. R. "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature", Geoscientific Model Development, 7(3), pp. 1247–1250, 2014. https://doi.org/10.5194/gmd-7-1247-2014

[33] Kvålseth, T. O. "Cautionary Note about R^2", The American Statistician, 39(4), pp. 279–285, 1985. https://doi.org/10.2307/2683704

[34] Python Software Foundation "Python, (3.10.0)", [computer program] Available at: https://www.python.org/downloads/release/python-3100/ [Accessed: 03 November 2025]

[35] GeoPandas Developers "GeoPandas, (0.14.0)", [computer program] Available at: https://geopandas.org/en/v0.14.0/ [Accessed: 03 November 2025]

[36] The SciPy Community "SciPy, (1.11.0)", [computer program] Available at: https://docs.scipy.org/doc/scipy-1.11.0/release/1.11.0-notes.html [Accessed: 03 November 2025]

[37] The Matplotlib Development Team "Matplotlib, (3.7.1)", [computer program] Available at: https://matplotlib.org/3.7.1/ [Accessed: 03 November 2025]

[38] Ollama "Ollama, (0.1.32)", [computer program] Available at: https://github.com/ollama/ollama/releases?page=19#release-v0.1.32 [Accessed: 03 November 2025]