Vision-Language Models for Structural Exposure Modeling from Street-Level Imagery

Yao Sun[1], Ahmed Abdelsalam[1], Xizhe Xue[2], Patrick Aravena Pelizari[3], Christian Geiß[3,4]

[1] German Aerospace Center (DLR), Remote Sensing Technology Institute (IMF), Weßling, 82234, Germany

[2] Technical University of Munich, Data Science in Earth Observation, Munich, 80333, Germany

[3] German Aerospace Center (DLR), German Remote Sensing Data Center (DFD), Weßling, 82234, Germany

[4] University of Bonn, Department of Geography, Bonn, 53115, Germany

Detailed information on building attributes, such as construction materials and structural types, is a fundamental prerequisite for accurate natural hazard risk assessment. Recent deep learning approaches based on convolutional neural networks (CNNs) have demonstrated the effectiveness of extracting such exposure-related information from street-level imagery, establishing a solid foundation for data-driven building characterization.

This study is motivated by the emerging capabilities of vision-language models (VLMs), which leverage large-scale pretraining and generalized visual-semantic reasoning to provide a unified framework for interpreting complex urban scenes. To assess their effectiveness in structural exposure modeling, we conducted comparative experiments using zero-shot inference and fine-tuning strategies. The dataset consists of over 29,000 annotated street-level façade images from the earthquake-prone region of Santiago, Chile.

The zero-shot results indicate that general-purpose off-the-shelf VLMs (e.g., InternVL2-8B) struggle to accurately infer complex structural engineering attributes due to insufficient domain-specific knowledge. In contrast, fine-tuning based on InternVL3-2B yields a substantial performance improvement: the model achieves high accuracy in building height estimation (90.6%) and roof shape classification (87.0%), and demonstrates strong performance in predicting lateral load-resisting system materials (78.8%) and complex seismic building structural types (SBST, 72.6%). These results suggest that fine-tuned VLMs can effectively acquire domain expertise, enabling scalable and low-cost exposure modeling. Future work will further investigate the potential of VLMs to infer latent structural characteristics through semantic reasoning.

Keywords: Exposure Modeling, Vision-Language Models (VLMs), Fine-tuning, Street-Level Imagery, Seismic Vulnerability.