An Evaluation of Zero-shot Classification of Multi-Modal Foundation Models for Point Clouds

Nils Sören Krause[1], Miguel Guzman-Merino[1] and Dennis Grauer[1]
[1]AI for sustainable construction, University of Rostock, Rostock, Germany
E-mail(s): nils.krause@uni-rostock.de, miguel.merino@uni-rostock.de, dennis.grauer@uni-rostock.de

Abstract: This paper investigates the use of various Multi-Modal Foundation Models (MMFMs) for classifying isometric views of point clouds. It addresses the challenge of reducing the amount of training data required for classification. MMFMs are prompted via their APIs using 2D-views from point cloud data. The paper examines the number of views needed to achieve efficient and accurate object classification in point clouds. We achieve an accuracy of up to 66%, and an F1-score of 35% is reached without specific training on the benchmark dataset. The results provide initial insights into the selection of MMFMs for zero-shot semantic segmentation of point cloud data. The approach is validated on parts of the ModelNet40 benchmark dataset.

Keywords: Multi-Modal Foundation Models, Classification, Zero-shot, Point cloud

DOI: 10.18154/RWTH-CONV-254887. Published in the conference proceedings of the 36. Forum Bauinformatik 2025, Aachen, Germany,

1 Introduction

As a step towards new labeled datasets in civil engineering, this paper examines the potential of Multi-Modal Foundation Models (MMFMs) in zero-shot classification of point clouds. It focuses on the MMFMs GPT-4o and Gemini-1.5-pro, as well as a reversed Stable Diffusion (Stable Diffusion-v1-5) model, in a comparative analysis. The test data is obtained from the ModelNet40 benchmark dataset [1]. The issue under discussion is whether the models can generate value labels for pre-classifying a dataset. Furthermore, the question arises as to whether the models are helpful for the semantic segmentation of the data. It has been demonstrated that these algorithms are capable of acquiring knowledge in a manner analogous to that of a human, through the utilization of previously classified data [2], [3]. The preparation of this classified data for the machine by a human is a laborious process. To reduce the effort required for semantic segmentation of point clouds, we are testing the behavior of classifications using Multi-Modal Foundation Models.

The paper begins with a State-of-the-art section, followed by the methodology and a demonstration of the tests. Afterwards, the test results and a discussion follow. Finally, we provide an outlook on further areas where research is needed. The results of this study help us to understand how to use MMFMs for point cloud classification.

The main contributions of the paper are:

- Evaluating GPT-4o, Gemini-1.5-pro and Stable Diffusion v1.5 for classifying isometric views of point clouds.
- Testing of different strategies for prompting the MMFM.
- Testing of different numbers of isometric views for classification.
- Classification without pre-trained models.
- Testing with a Benchmark dataset to ensure reproducible results.

2 State-of-the-art

Since the introduction of ChatGPT in 2022, the development and use of MMFMs for various applications have accelerated rapidly. Models such as GPT-4o from OpenAI, Gemini from Google, Claude from Anthropic, or DeepSeek demonstrate that they can process text, images, videos, and code. While they have shown potential in generative tasks–such as creating architectural images [4], floor plans [5], or assisting in the building design phase [6]–the understanding of point clouds remains underexplored. The most significant challenge is analyzing the unstructured data derived from LiDAR scans, for which deep learning techniques are employed [7]. There are still differences in the methods to classify point clouds; some approaches use raw point clouds [8], [9] for the segmentation and classification of point clouds. However, these approaches can only recognize data with which they are familiar. The performance of these models depends heavily on specific parameters that must be adapted to the models. They also require extensive datasets for training or testing of these models, as demonstrated by [10]. There are still datasets for point clouds [11], [12]. However, it leaks labeled datasets, particularly in civil engineering. To obtain these datasets for civil engineering, we need to examine how to label the point cloud data with minimal manual effort. For the point cloud pre-processing task, the State-of-the-art is also looking for new direction, they are going to use large language models for the point cloud alignment [13]. Based on these, we will take a deeper look at Prompt variation and the capabilities of MMFMs for image recognition of point clouds. To use MMFMs for point cloud understanding, we transform the point cloud data into images by generating isometric views. The works of [14] and [15] enable us to classify point clouds in the form of images, also named zero-shot classification. For example, the work of [16] presents a model approach named PointBLIP for zero-shot classification of point clouds, tested on ModelNet40, with an accuracy of 66.25 %.

3 Methodical

The 3D-data originates from 3D-mesh objects collected as OFF files in the ModelNet40 dataset [1]. In step (i), the ModelNet40 dataset was downloaded. The ModelNet40 dataset consists of 40 classes that can be utilized for semantic segmentation. In step (ii), the OFF objects contained within the data set were converted into point clouds. The point clouds consist of area-weighted distributions with exactly 3000 points. In step (iii), one to four isometric views were extracted at the following angles: *front, side, top, and diagonal*. Images contain a white background and black dots, sized at 512×512 pixels, exported in PNG-format. In step (iv), these views were then prompted into the various MMFMs via an API connection. In step (v), ensure its alignment with the established ground-truth for effective

classification based on ModelNet40 labels. The entire code and used labels for data preparation and evaluation can be viewed in our GitHub [1]. In Figure 1, you can see an example of the point cloud from the used dataset ModelNet40. We did not use the complete dataset because the API costs for prompting were too high.

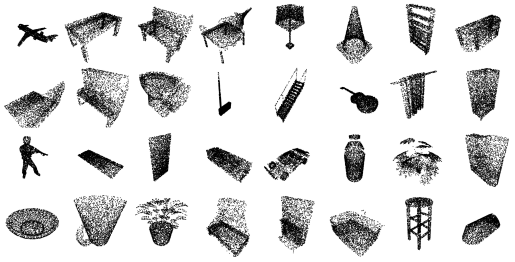

Figure 1: Isometric views of the point clouds from the used ModelNet40 dataset.

3.1 Multi-Modal Models

For the evaluation, several current MMFMs were selected. Based on their ability to process visual and textual input via their APIs, priority was given to models with public availability. Stable Diffusion was chosen due to its captioning capability, allowing for a comparison of the visual understanding of different models. The following models in Table 1 were selected:

Table 1: Overview of tested MMFMs.

Model	From	API	Multi-Modal skills
GPT-4o	OpenAI	Yes	Image + Text
Gemini-1.5-pro	Google	Yes	Image + Text
Stable Diffusion	Stable Diffusion-v1-5	local	Image + Text(encoder)

3.2 How to test

The following prompt in Algorithm 1 was used for evaluation. The prompt only varies in the words: *"three isometric"* for the respective number of isometric views:

Algorithm 1: Prompt for isometric classification.

```
prompt = (
You are shown three isometric views of a 3D object. Choose exactly one of the following
    words that best describes the
object based on the three isometric views. Reply with only one word from this list (no
    extra text):
join(sorted(list(class_list)))
```

The ground-truth data is based on the labels of ModelNet40. The individual classes were transferred to the MMFMs as a class list, which is shown in Algorithm 2, so that they do not deviate from the ground truth.

Algorithm 2: Class list for the prompt.

[1] https://github.com/AI4SC/Multimodal-point-cloud-classifier.git

```
class_list = {
"airplane", "bed", "bench", "bookshelf", "bottle", "bowl", "car", "chair", "cone", "cup",
    "curtain", "door", "flower_pot",
"glass_box", "guitar", "keyboard", "lamp", "laptop", "mantel", "person", "piano", "plant"
    , "radio", "range_hood",
"sink", "stairs", "stool", "tent", "toilet", "tv_stand", "vase", "wardrobe", "xbox"}
```

The evaluation is then carried out using the following metrics:

- **Accuracy**: The accuracy is the overall performance of the true predicted labels by the different models. The accuracy is expressed as a percentage of true predicted labels.
- **F1-Score**: Averages the F1-score for each class without considering class imbalances, known as the macro-average F1-Score.

3.3 Real data testing

We tested the workflow using synthetic data and then applied it to real data scanned with the BLK2GO from Leica. The data was segmented from the point cloud using CloudCompare. We generated a single labeled point cloud for each object in a class. The scene contains a 3D-printer, beamer, desktop, cardboard box windows, heater, lamp, shelf, robot arm, robot dog, chair, and table. The key difference lies in our ability to extract geometric features from the point cloud, including height, brightness, width, and color. The geometric features are derived from an axis-aligned bounding box (AABB) using Opend3D with Python. The scene is shown in Figure 2. The dataset for testing was modified to ensure that the objects have their real dimensions. For the evaluation of the influence of geometric features, the analysis is conducted, including consideration of these features. Due to the effort required for segmentation, only 35 objects that are present in the scene were tested.

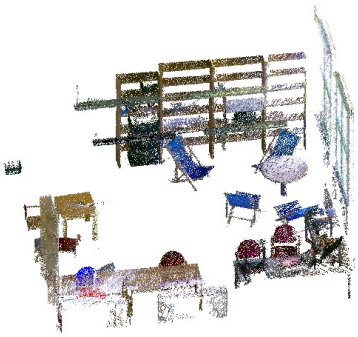

Figure 2: Tested scene of the scanned real data, captured with the BLK2GO

The class list in Algorithm 2 was updated to reflect the new objects. The prompt for testing the models is shown in Algorithm 3. The objects were tested based on four isometric views with the different models.

Algorithm 3: Prompt for isometric classification with AABB.

```
prompt = (
```

```
You are shown four isometric views of a 3D object. The object has the following geometric
    properties
(note: the height, width, and depth are derived from the objects axis-aligned bounding
    box, and may
not represent precise physical dimensions but give a general sense of size and
    proportions)
Height: features['height']
Width: features['width']
Depth: features['depth']
Based on the isometric views and the geometric dimensions, choose exactly one word from
    the list below that best
describes the object. Respond with only one word from this list: join(sorted(class_list))
```

4 Results

The evaluation of the data is presented in Figure 3 and Table 2, which includes the model, samples, accuracy, and F1-score. We evaluated the classification capability of the model GPT-4o and Gemini-1.5-pro, as well as a reversed Stable Diffusion model with a captioning function. The results show that the MMFM Gemini-1.5-pro has better results in every validation scenario. Overall, the classification was even more precise with colored point clouds and the addition of geometric features. The results indicate that the captioning function in Stable Diffusion can be used to generate a partial description of the classification of one isometric view of the point clouds. Filters were used for captioning processing, which not only searched for the required class label but also for synonyms similar to the class. The model achieved an accuracy of 17 %.

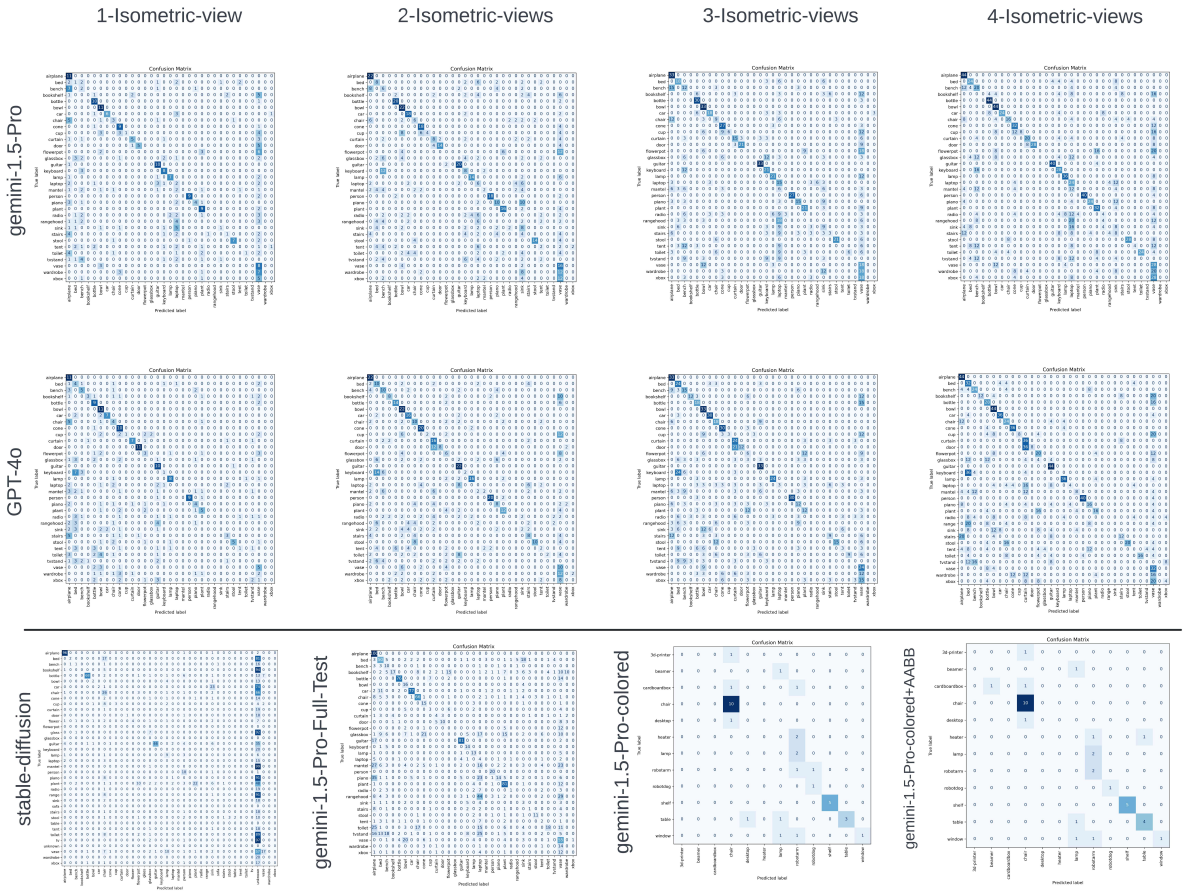

Figure 3: Confusion Matrix's of the classification results, from the evaluated models(First row: Gemini-1.5-pro, 1-4 Isometric views; Second row: GPT-4o, 1-4 Isometric views; Third row: Stable Diffusion: ModelNet40, test data big test, Gemini-1.5-pro-big-test: ModelNet40, big test, Gemini-1.5-pro-colored: colored real point cloud BLK2GO, Gemini-1.5-pro-colored: colored real point cloud BLK2GO + AABB).

Table 2: Classification accuracy and F1-score of GPT-4o and Gemini-1.5-pro on isometric views and different evaluations.

Model Variant	**Samples**	**Accuracy**	**F1-score**
GPT-4o-1-Iso.-view	363	0.38	0.32
GPT-4o-2-Iso.-views	726	0.39	0.33
GPT-4o-3-Iso.-views	1089	0.40	0.34
GPT-4o-4-Iso.-views	1452	0.41	0.35
Gemini-1.5-pro-1-Iso.-view	363	0.36	0.32
Gemini-1.5-pro-2-Iso.-views	726	0.39	0.35
Gemini-1.5-pro-3-Iso.-views	1089	0.40	0.36
Gemini-1.5-pro-4-Iso.-views	1452	0.43	0.39
Gemini-1.5-pro-1-Iso.-view-big-test	1861	0.41	0.32
Stable-Diffusion-1-Iso.-view; CLIP/ViT-L-14/openai-big-test	1861	0.17	0.14
GPT-4o-4-Iso.-views-colored-cloud	35	0.54	0.29
GPT-4o-4-Iso.-views-feature-geometries	35	0.57	0.29
Gemini-1.5-pro-4-Iso.-views-colored-cloud	35	0.60	0.33
Gemini-1.5-pro-4-Iso.-views-feature-geometries	35	0.66	0.35

5 Discussion and outlook

The evaluation results presented in Table 2 and Figure 3 provide insights into the classification capabilities of two MMFMs (GPT-4o and Gemini-1.5-pro). In initial tests, the performance of the classification improves with the increasing number of isometric views provided. For example, the GPT-4o model increases from 38 % to an accuracy of 41 %. In contrast, the Gemini-1.5-pro model increased the accuracy from 36 % to 43 %. These results confirm that the classification with multi-view input of point clouds has a positive influence. For Future development, we approach to test the use of more than four isometric views. Beyond the visual input, we tested the benefits of enhancing the prompt with additional features from the point cloud. Classification performance increases if the point cloud is colored. Otherwise, the implementation of geometric features from AABB as textual input to the model demonstrates an understanding of models for context information of objects. For example, the accuracy in classification for tables increases with the known possible geometric dimension of the object. But these need to be evaluated in further studies. The Stable Diffusion model achieves an accuracy of 17 % in captioning images. Stable Diffusion accuracy is 25 % lower than the MMFM results from the Gemini-1.5-pro test. A closer look at the results from the confusion matrix in Figure 3 shows differences in the classification results of the objects. We can observe that, for example, the GPT-4o model struggles to recognize the following objects: glass boxes, keyboards, radios, range hoods, sinks, tents, TV stands, and wardrobes. The Gemini-1.5-pro model has issues with the following objects: bookshelves, flowerpots, glass boxes, mantels, sinks, stairs, TV stands, wardrobes, and Xboxes. But the Gemini-1.5-pro model was able to recognize these objects with a higher hit rate, whereas GPT-4o had problems with radios, tents, and keyboards. Both models had difficulties with these objects: glass boxes, sinks, TV stands, and wardrobes. Now, look at the geometric properties in Figure 1. For example, whether it is a rectangular object or a circular object, it becomes opaque. The objects that the model has recognized incorrectly are only partly similar in their geometric properties. We can recognize this clearly with the confusion matrix, which is based on the predicted labels and the actual class of the vase. Both models accurately labeled a wide range of objects in that column, but they struggled with mostly rectangular structures. Below the diagonal of the confusion matrix, Gemini shows a slight cluster of misclassified labels between the airplane and flowerpot classes. In contrast, GPT-4o distribution is spread across the airplane, up to the bench and laptop classes. Furthermore, objects such as planes, guitars, people, and chairs are easily recognized by the models, as they stand out clearly from other objects that usually consist only of a rectangular or circular surface. This suggests the model lacks by details to recognize the objects. From the distribution of the other objects in the confusion matrix. The Stable Diffusion model shows that it is rather intended for the generation of labels for images of point clouds. For future work, we tested the semantic segmentation for construction site objects. Figure 4 shows the results of the outlook.

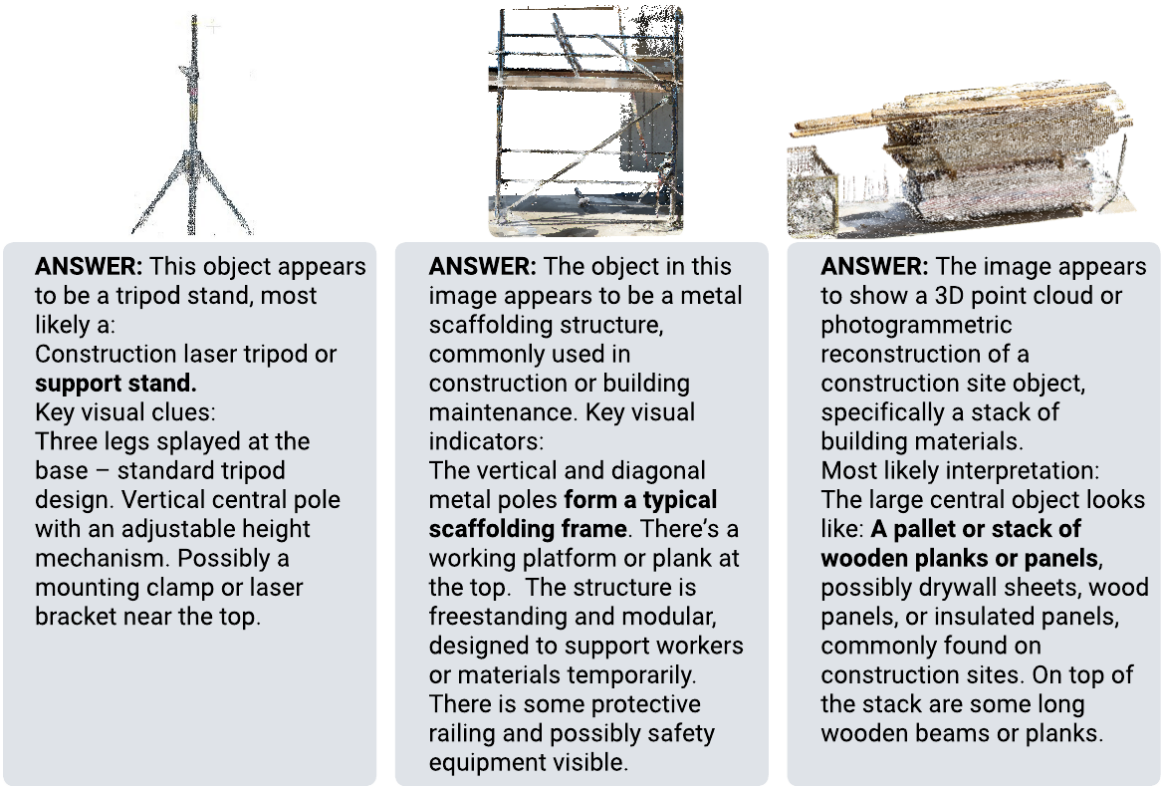

Figure 4: Outlook of GPT-4o-Model with point cloud data from a construction site. The figure shows the answers of the model after prompting the model with the question: What is it?

6 Conclusion

The results show that available MMFMs are alternative tools for the pre-classification of new datasets. These datasets are important because well-segmented and semantically segmented point cloud data are not widely available in the construction sector. This work represents a step toward new methods and capabilities for obtaining labeled point cloud data more easily and quickly by using MMFMs. But they still have limitations in terms of accuracy or in the prompting chain of the model. The approach can benefit from chain-of-thought reasoning and expanding the geometric information, e.g., with density histograms or eigenvalues based on shape ratios. But this is a point for future research. The comparison of the MMFMs used with other pre-trained deep learning approaches for point cloud classification on ModelNet40 does not come close to these, in the range of 66 % to 99 % [16] [2]. Our evaluation reached a maximum of 43 % with ModelNet40 and 66 % with real data.

References

[1] Z. Wu et al., "3d shapenets: A deep representation for volumetric shapes", in *IEEE conference on computer vision and pattern recognition*, 2015.

[2] H. Zhang et al., "Deep learning-based 3d point cloud classification: A systematic survey and outlook", *Displays*, 2023.

[2] https://modelnet.cs.princeton.edu/

[3] E. Grilli, F. Menna, and F. Remondino, “A review of point clouds segmentation and classification algorithms”, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2017.

[4] V. Paananen, J. Oppenlaender, and A. Visuri, “Using text-to-image generation for architectural design ideation”, *Journal of Architectural Computing*, 2024.

[5] J. Ploennigs and M. Berger, “Automating computational design with generative ai”, *Civil Engineering Design*, 2024.

[6] C. Zhong, L. Yi’an Shi, and L. Wang, “Ai-enhanced performative building design optimisation and exploration”, in *29 Conference on Computer-Aided Architectural Design Research*, 2024.

[7] S. A. Bello, S. Yu, C. Wang, J. M. Adam, and J. Li, “Deep learning on 3d point clouds”, *Remote Sensing*, 2020.

[8] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation”, in *IEEE conference on computer vision and pattern recognition*, 2017.

[9] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “Pointnet++: Deep hierarchical feature learning on point sets in a metric space”, *Advances in neural information processing systems*, 2017.

[10] G. Qian et al., “Pointnext: Revisiting pointnet++ with improved training and scaling strategies”, *Advances in neural information processing systems*, 2022.

[11] M. De Deuge, A. Quadros, C. Hung, and B. Douillard, “Unsupervised feature learning for classification of outdoor 3d scans”, in *Australasian conference on robitics and automation*, 2013.

[12] M. A. Uy, Q.-H. Pham, B.-S. Hua, T. Nguyen, and S.-K. Yeung, “Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data”, in *CVF international conference on computer vision*, 2019.

[13] Z. Guo et al., “Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following”, *arXiv:2309.00615*, 2023.

[14] N. Zhao, T.-S. Chua, and G. H. Lee, “Few-shot 3d point cloud semantic segmentation”, in *IEEE/CVF conference on computer vision and pattern recognition*, 2021.

[15] R. Zhang et al., “Pointclip: Point cloud understanding by clip”, in *IEEE/CVF conference on computer vision and pattern recognition*, 2022.

[16] Y. Xiao, Y. Dou, and S. Yang, “Pointblip: Zero-training point cloud classification network based on blip-2 model”, *Remote Sensing*, 2024.