Information Fusion 129 (2026) 104007

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

GLUE3D: General language understanding evaluation for 3D point clouds

Giorgio Mariani *, Alessandro Raganato, Simone Melzi, Gabriella Pasi

Universitá degli Studi di Milano-Bicocca, Piazza dell'Ateneo Nuovo 1, Milan, 20126, Italy

ARTICLE INFO

Keywords:
Multimodal large language models
Point cloud
3D understanding
Benchmark

ABSTRACT

Multimodal Large Language Models have achieved impressive results on text and image benchmarks, yet their capacity to ground language in 3D geometry is still largely unexplored. Existing 3D evaluations are either confined to specialised domains, such as indoor scans, or hampered by poor texture fidelity, and none allow a fair, modality-aligned comparison with the 2D counterparts. Without a rigorous benchmark, it remains unclear whether current 3D-aware models genuinely grasp shape, colour, pose, and quantity, or merely echo memorised textual priors.

We address this gap with GLUE3D (*General Language Understanding Evaluation for 3D Point Clouds*), a benchmark built around 128 richly textured meshes spanning creatures, objects, architecture and transport. Each asset is provided both as a 50 k-point RGB point cloud and as a matched 512×512 rendering, enabling point-for-point evaluation across modalities. Over these assets we manually curate 1024 binary probes, 256 multiple-choice questions, 256 open-ended questions, and 128 caption prompts that jointly evaluate sub-entity recognition, physical state, colour attribution and counting, giving a fine-grained understanding of 3D geometry and its semantics. A comprehensive study of twelve recent systems reveals a pronounced modality gap: the image-conditioned Qwen-2.5-VL achieves 79 % accuracy on binary probes and 74 % on multiple-choice questions, while the best point-cloud model reaches only 55 % and 33 % respectively. Caption-quality assessments follow the same pattern, underscoring the gap that remains for genuine 3D understanding. We make GLUE3D publicly available, along with its evaluation scripts and baseline scores, to advance progress in geometry-aware multimodal language understanding.

1. Introduction

Recent advances in artificial intelligence have led to the development of *Multimodal Large Language Models* (MLLMs). MLLMs are able to integrate and understand multiple data modalities—such as text, images, and 3D geometry—to perform tasks that would be otherwise impossible or difficult to approach, like image captioning, visual question answering, image-conditioned generation, and many more. These models typically function by aligning modality-specific encoders with a language backbone, often an instruction-tuned large language model. While MLLMs have shown impressive capabilities, the integration of different modalities introduces new challenges. For example, misalignment between the encoder and the language model can lead to hallucinated or incorrect outputs, where the model relies on learned textual priors rather than grounded perceptual understanding [1]. This highlights the need for comprehensive evaluation methodologies for assessing MLLM performance, especially when compared to the more mature unimodal benchmarks. In this work, we focus on a specific subclass of MLLMs: *3D Large Language Models* (3D-LLMs), which are designed to understand and reason over three-dimensional inputs and connect them to language. These models enable tasks such as scene description, object-centric reasoning, and spatial understanding beyond what 2D images can typically support. Among the various 3D representations, we adopt point clouds as our modality of interest due to their widespread use in domains such as autonomous driving, robotics, and augmented reality. Point clouds provide rich spatial detail while remaining computationally efficient for large-scale data processing.

Existing benchmarks. While numerous 2D multimodal benchmarks have been developed to assess and diagnose vision-language models [2–6], analogous resources in the 3D domain remain limited. Existing 3D datasets often focus on narrow categories (e.g., indoor scans or street-level reconstructions) and suffer from significant quality issues when point clouds are derived from 3D asset-sharing platforms (e.g., Sketchfab, CGTrader, Thingiverse). In particular, color fidelity is frequently

* Corresponding author.
E-mail addresses: giorgio.mariani@unimib.it (G. Mariani), alessandro.raganato@unimib.it (A. Raganato), simone.melzi@unimib.it (S. Melzi), gabriella.pasi@unimib.it (G. Pasi).

https://doi.org/10.1016/j.inffus.2025.104007
Received 30 June 2025; Received in revised form 6 November 2025; Accepted 25 November 2025
Available online 29 November 2025

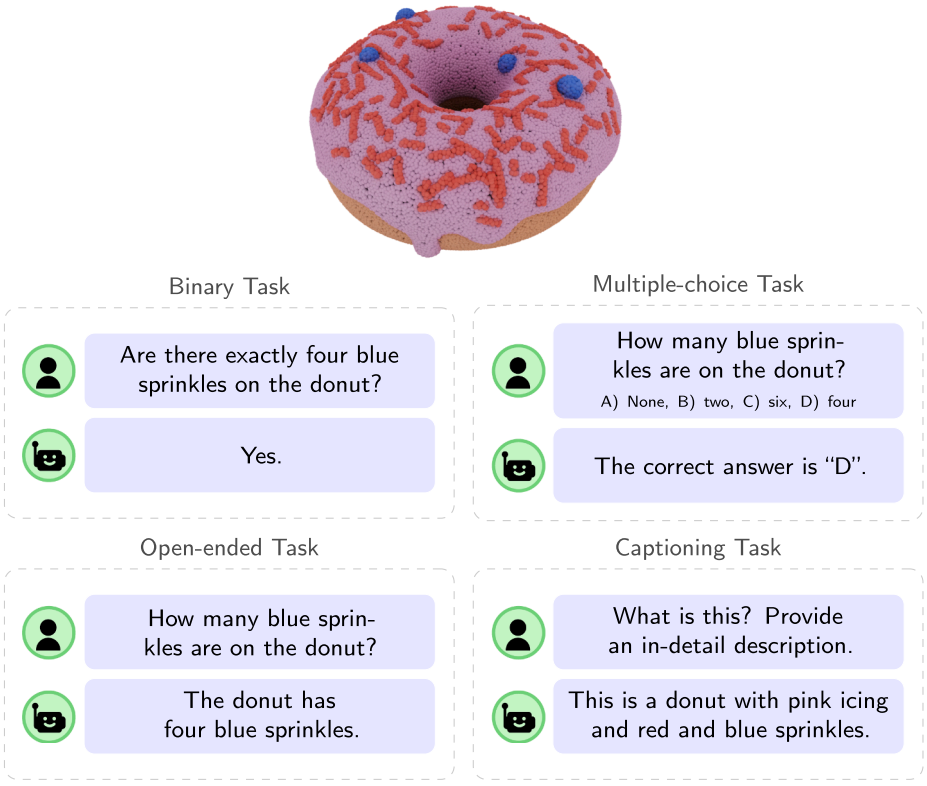

Fig. 1. Visual overview of the four evaluation tasks included in GLUE3D. The benchmark consists of four task types: binary question answering, multiple-choice question answering, open-ended question answering, and captioning. This diverse set of tasks enables a more robust and comprehensive assessment of multimodal understanding in 3D-LLMs.

degraded due to texture mispackaging and incompatibilities across proprietary 3D file formats, many of which require specific rendering software (e.g., Maya, Blender, Z-Brush) for correct interpretation. These limitations hinder the development and evaluation of models capable of robust 3D semantic understanding. Moreover, to the best of our knowledge, there exists no benchmark that enables a direct comparison between 2D and 3D LLMs. While prior work has explored each modality in isolation, a unified evaluation framework that allows for modality-aligned inputs-such as matched point clouds and rendered images-is currently missing.

Our contributions. Inspired by similar work in the vision domain [3,5,6], we construct a benchmark composed of 1024 binary, 256 multiple-choice, 256 open-ended, and 128 captioning questions related to both point cloud data and images. Fig. 1 provides an example of each of our question types on a 3D object representing a donut. Furthermore, for our binary task, each question is designed to probe a 3D-LLM's understanding of specific properties. Given an input point cloud, the model's understanding of color, state, number, and potential sub-components is assessed. The aim of our work is twofold: (i) to provide a carefully curated, high-quality benchmark[1] for evaluating the capabilities of 3D-LLMs; (ii) to shed light on the performance discrepancy between current state-of-the-art 2D-LLMs and comparable 3D-LLMs.

Paper structure. The paper is organized as follows: Section 2 introduces related works, i.e., similar benchmarks and/or datasets, both in the image and 3D domains. Section 3 provides an in-depth description of our benchmark, detailing the object categories and question types it includes, as well as the data collection and annotation processes. Finally, Sections 4 and 5 describe our experimental setup and our findings, respectively.

[1] The benchmark data and evaluation scripts are available at https://github.com/giorgio-mariani/GLUE3D.

2. Related work

2.1. Multimodal large language models

Large Language Models (LLMs) have shown notable capabilities in natural language understanding and generation [7]. However, many real-world tasks require the interaction between multiple modalities, such as vision, audio, and text. This led to the emergence of *Multimodal Large Language Models* (MLLMs), which extend traditional LLMs by processing and integrating information from diverse data sources.

Common MLLM architectures. MLLMs are typically built by aligning or integrating pretrained modality encoders (e.g., CLIP, ViT) with powerful LLM backbones (e.g., GPT, LLaMA, Vicuna). This design enables the model to process visual inputs such as images, videos, or 3D point clouds alongside text. The fusion of modalities allows MLLMs to perform a broad range of tasks, including visual question answering, image captioning, visual grounding, and embodied reasoning.

Despite their progress, MLLMs face several challenges. For example, ensuring that visual and textual representations are meaningfully integrated, that is, representation coming from the modality encoder must be properly aligned with the one found in the textual backbone, thus ensuring a common ground between the two models [1,2]. This misalignment, among other things, may cause the generation of inaccurate or fabricated content not grounded in visual input [8]. Furthermore, high-quality, large-scale multimodal datasets are often limited compared to unimodal corpora. Because of this, developing robust benchmarks that can meaningfully measure multimodal reasoning and understanding can sometimes prove difficult.

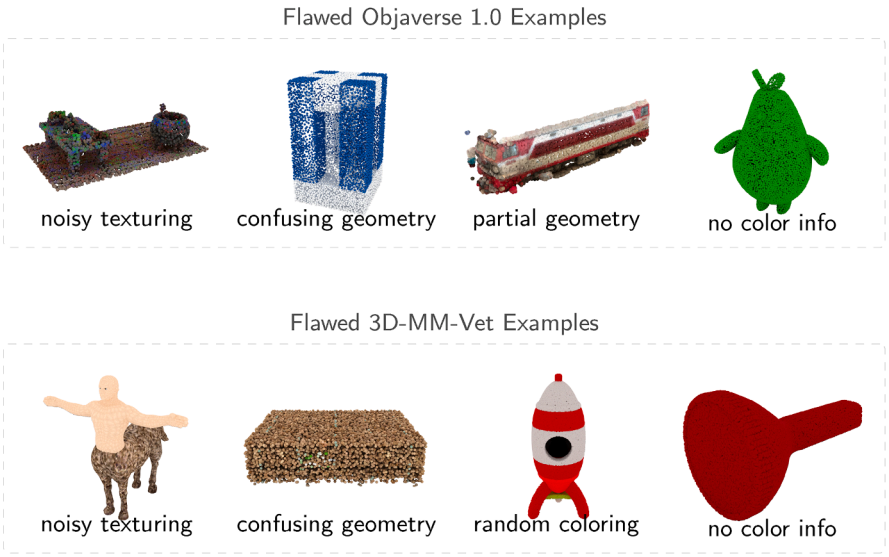

Fig. 2. Qualitative examples of various flaws in existing 3D-LLMs benchmarks. The origins of these issues are several, mostly caused by a combination of the automatic point cloud extraction process and inconsistent data information (such as incorrect texture paths) at the source.

It is therefore clear that the development of reliable benchmarks is essential to improving the accuracy and effectiveness of future MLLMs.

2.2. Evaluation methodologies for MLLMs

Two-dimensional domain. Traditional vision-language benchmarks have primarily targeted specific capabilities of interest, such as visual recognition [9], image captioning [10,11], scene text understanding [12,13], and several more. However, the emergence of general-purpose MLLMs has created a pressing need for updated benchmarks that reflect the complexity and integration demands of modern multimodal tasks.

This demand has led to the development of 2D-LLM-specific benchmarks designed to assess broad multimodal understanding across a variety of tasks. For instance, the *Multimodal Model Evaluation* (MME) benchmark [4] targets fine-grained visual perception and reasoning, using binary questions organized into 14 coarse- and fine-level categories over a curated image dataset. Similarly, *MMBench* [5] evaluates MLLMs on a diverse set of perception and reasoning tasks, and additionally introduces multilingual support in both English and Chinese. MMBench adopts a free-form-to-multiple-choice conversion strategy to standardize answer evaluation. The *Multimodal-Veterinarian* (MM-Vet) benchmark [6] takes a different approach by leveraging the LLM-as-a-Judge framework [14] to evaluate models on tasks that require multiple overlapping competencies, such as spatial awareness, factual knowledge, OCR capabilities, and more. In parallel, more specialized benchmarks have been proposed to target specific issues such as hallucination. The *Polling Object Probing Evaluation* (POPE) benchmark [3], for example, diagnoses hallucination tendencies in 2D-LLMs using binary probing questions constructed from MSCOCO annotations [10].

Three-dimensional domain. Arguably, the development of 3D-LLMs is still in its early stages, and accordingly, the field lacks well-established benchmarks for comprehensive evaluation. Existing benchmarks are often limited in scope or data quality. For example, 3D-POPE [15] focuses specifically on evaluating object hallucinations for indoor scenes, providing valuable insights but within a narrow task domain. Other benchmarks, such as those proposed in [16,17], are built on datasets with often suboptimal texture quality, which can compromise the representativeness and generalizability of their evaluation results. Furthermore, they both utilize GPT-as-a-Judge strategies, making the evaluation less long-lasting and depending on the current OpenAI available models. As such, there remains a need for high-quality, diverse, and task-rich benchmarks to effectively assess the capabilities of 3D-LLMs.

Table 1
Comparison between GLUE3D and existing 3D object understanding benchmarks. The percentages indicates the number of surfaces in each benchmark that have a singular—or very few—colors in the point cloud (all colors are quantized into 8^3 bins). Furthermore, we can see that GLUE3D offers a greater amount of questions, arguably providing an overall better evaluation of fine-grained understanding capabilities of 2D and 3D-LLMs.

Benchmark	Shape number	Question number
3D-MM-Vet	59 shapes	232 questions
PointLLM test set	200 shapes	200×3 questions
GLUE3D	128 shapes	$1024 + 256 \times 2 + 128$ questions
	Single color	Five colors ($\leq$)
3D-MM-Vet	14.40 %	35.5 %
PointLLM test set	12.0 %	15.5 %
GLUE3D	0.00 %	0.00 %

3. The GLUE3D benchmark

In this section, we introduce our benchmark and describe its core design principles and methodology. Following established practices in multimodal evaluation [3,18], our framework primarily relies on binary questions to assess the perceptual and semantic understanding capabilities of MLLMs. The full benchmark comprises 1024 binary questions, 256 multiple-choice questions, 256 open-ended questions, and 128 captioning instances.

2D vs 3D. One of our aims for this benchmark is to assess the existence—and extent—of performance disparities between publicly available 2D-LLMs against common 3D-LLMs. To enable such a comparison, each three-dimensional asset in our benchmark is accompanied by human-curated two-dimensional renderings. This dual-modality setup allows for a controlled evaluation of point cloud-based models alongside image-based vision-language models. Specifically, the benchmark includes 128 high-quality textured polygon meshes, from which we derive three distinct representations: a 3D point cloud, a single 2D image, and multiple-view images of the same object. Each point cloud comprises 50,000 points sampled from the mesh surface, capturing

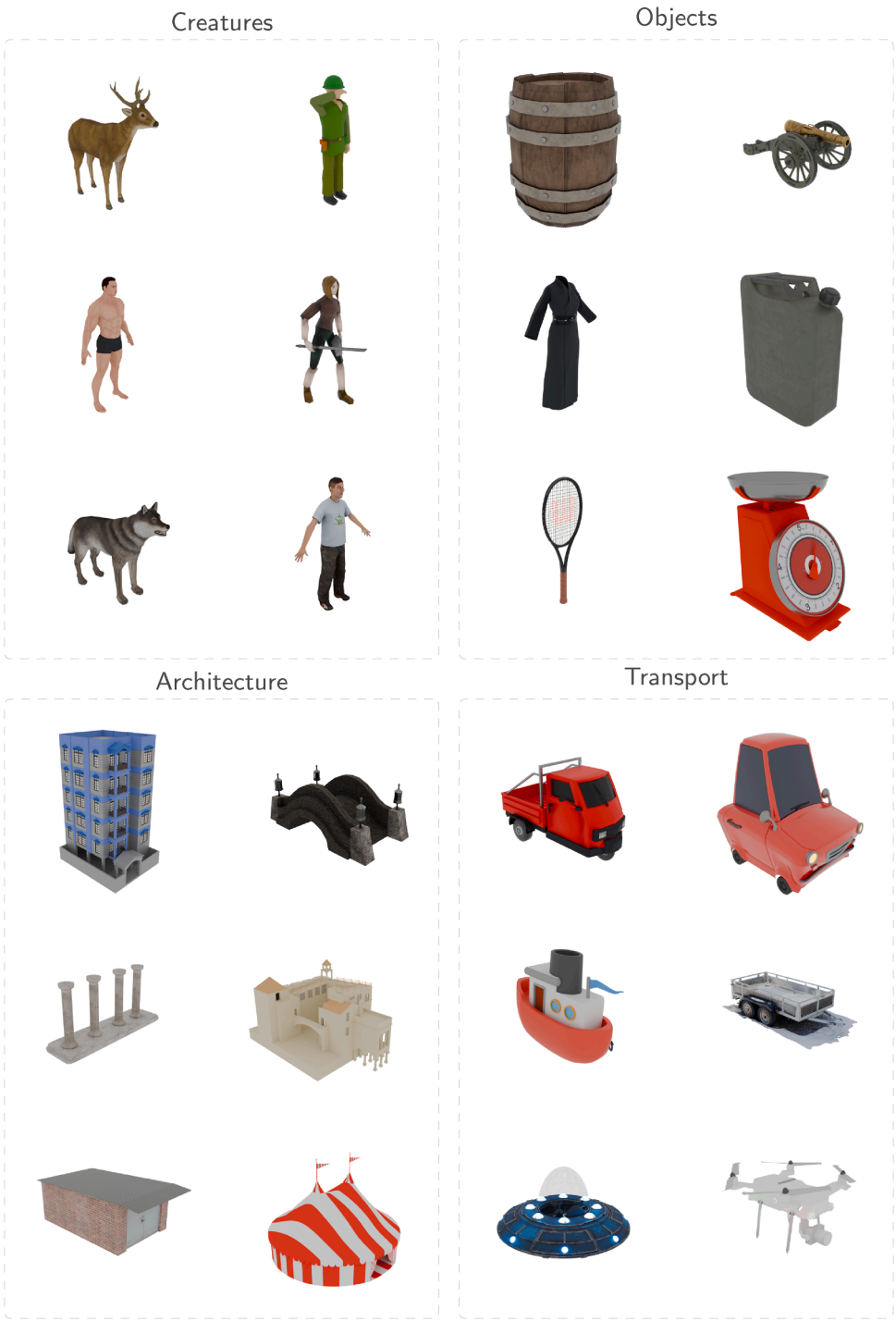

Fig. 3. Representative subsamples of GLUE3D 3D-surfaces. The 3D geometries offered by GLUE3D are both high-quality and diverse in style and coloring, thus providing a strong foundation for our evaluation tasks.

fine-grainedgeometric detail. Corresponding 2D images are rendered using a human-provided camera view, at a resolution of 512×512 RGB pixels (each channel ranging from 0 to 255). Bridging between the 2D and 3D domains, we adopt a multiple-view setting, in which five RGB images (optionally accompanied by depth maps and camera parameters) are utilized instead of a single image or a single RGB point cloud. Ideally, this setting should provide more information than a single image, though at the cost of increased computational requirements due to the use of several images during inference. More information on the point cloud and rendering generation procedure can be found in Section 3.3.

3D Geometry. The 128 surfaces contained in our benchmark can be categorized into four distinct classes—*creatures*, *objects*, *architecture*, and *transport*—each containing exactly 32 instances. This classification encompasses object types commonly encountered in computer graphics and 3D processing scenarios, ensuring broad applicability across vari-

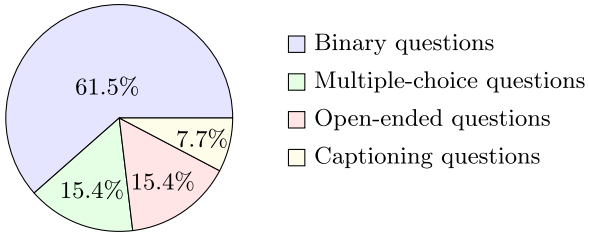

Fig. 4. Distribution of question types in GLUE3D. The benchmark includes 1024 binary question-answer pairs, 256 multiple-choice question-answer pairs, 256 open-ended question-answer pairs, and 128 captioning question-answer pairs.

ous domains. For each of these surfaces, an average of eight binary, two multiple-choice, and one captioning question-answer pairs are provided.

Comparison with existing 3D benchmarks. As discussed in Section 2, multiple benchmarks for evaluating 3D-LLM object understanding have already been introduced [16,17]. However, they face some notable limitations: the quality of point cloud color annotations, the overall benchmark size (e.g., 3D-MM-Vet), and fewer fine-grained object understanding questions.[2] Moreover, both benchmarks only provide raw point clouds, which restricts the generalization to different architectures operating on different 2D/3D representations. In contrast, our benchmark makes textured meshes available, enabling a possible expansion through variations in lighting conditions and/or surface representations (e.g., signed distance functions, Gaussian splats). We argue that this ability to generalize to different contexts allows for a more future-proof and universal benchmark, reducing the limitations of a specific 3D representation.

As we mentioned, both benchmarks suffer from the presence of several poor-quality point clouds, mostly due to the automatic point cloud extraction procedure and errors (or inconsistencies) in the original three-dimensional data. Some representative examples of common issues that can be found are shown in Fig. 2. These can be broadly categorized into the following: (i) *Missing color information.* The original mesh may lack texture data; this may be because the original mesh does not provide it, or because the texture paths in the mesh are incorrect (or assume a specific folder structure). (ii) *Low quality surfaces.* Some meshes may be lower-quality and not easily understandable even to a human observer. The inclusion of these in a benchmark may lower its evaluation capabilities. (iii) *Incorrect rendering setup.* Some surfaces may require specific rendering setups and/or rendering engines. Using the default Blender rendering setup may produce artifacts and incorrect color information. Table 1 shows the percentage of surfaces with no color information (i.e., a single color), and/or very limited color information. As can be seen, both the PointLLM test set and 3D-MM-Vet contain a significant amount of low color-information point clouds. We argue that this may hinder their overall quality (Fig. 3).

3.1. Benchmark tasks

Inspired by prior work in the field [3,4,6,16,17], our benchmark integrates a diverse set of evaluation methodologies to comprehensively assess model capabilities. First, we adopt a binary question format to probe the model's understanding of specific properties, following methodologies established in [3,4]. In addition, we include a set of more challenging multiple-choice questions, inspired by formats used in [5], which may require more complex reasoning capabilities. Finally, we incorporate both an open-ended variant of the multiple-choice questions and an open-ended captioning task, wherein the model is asked to generate descriptive summaries of 3D shapes. These generated captions are then evaluated against human-written references using a unimodal LLM-based scoring system, similar to the approach taken in [16,17].

By combining these complementary evaluation strategies, our benchmark provides a more holistic and interpretable assessment of multimodal LLM performance compared to current benchmarks in the 3D domain [15–17].

Binary Questions. As illustrated in Fig. 4, the majority of GLUE3D consists of binary questions, which serve as focused probes into the comprehension capabilities of the MLLMs under evaluation. To enable a fine-grained analysis of model performance and failure modes, we categorize these binary questions into four distinct types, each targeting a specific dimension of 3D semantic understanding: (i) *subentity understanding*, (ii) *pose or physical state recognition*, (iii) *color attribution*, and (iv) *quantity estimation*. This differentiation facilitates a more structured evaluation of robustness and generalization.

- *Subentity understanding:* 3D assets in our benchmark often represent complex composite entities—such as characters, vehicles, or architectural structures—that consist of multiple semantically meaningful parts. Subentity questions probe whether the model correctly identifies which constituent components are plausibly present or hallucinates plausible but absent elements (e.g., assuming a sword in the hand of a medieval character or a pilot in an aircraft).
- *Pose/State recognition:* Certain objects or subentities admit a set of physically plausible configurations. Pose-related questions assess the model's ability to reason about such states; for instance, identifying whether a tire appears deflated or whether a hand is open or closed. Errors in this category may suggest difficulties in interpreting fine-grained geometric variation.
- *Color attribution:* Although the point clouds are colorized using RGB values, models may still hallucinate canonical or expected colorations based on learned priors rather than grounded visual input. Alternatively, models may misattribute color to the wrong segment (e.g., inverting the color of a vehicle's roof and hood), indicating a lack of spatial-color alignment.
- *Quantity estimation:* Counting subcomponents or repeated elements remains a difficult challenge for many multimodal LLMs [2]. Questions in this category test whether the model can accurately estimate the number of specific parts (e.g., wheels on a car or blades on a fan), providing insight into its capacity for global and local structure parsing.

By designing questions along these four axes, GLUE3D enables a more targeted and interpretable assessment of where and how 3D-LLMs may succeed or fail. Examples for each question type can be seen in Fig. 5.

Multiple-Choice Questions. To complement the binary question framework, we include a smaller but more complex set of multiple-choice questions aimed at evaluating higher-level semantic reasoning in 3D-LLMs. Unlike binary questions that test the presence or absence of specific properties, multiple-choice questions require models to select the most accurate answer from a set of plausible alternatives.

Open-ended Questions. We further introduce an open-ended variant of our multiple-choice task. These questions are derived directly from the multiple-choice set (often the same) but allow the model to freely generate its response rather than select from predefined options. Each open-ended question-answer pair was manually converted from its corresponding multiple-choice counterpart to preserve semantic consistency.

[2] The PointLLM test set only has broad object understanding tasks: a closed-vocabulary object classification task, an open-vocabulary object classification task, and finally a captioning task. 3D-MM-Vet offers more complex questions, probing for different understanding capabilities, however fine-grained visual recognition question account for only 59 questions (i.e., $\sim 25\,\%$ of the benchmark).

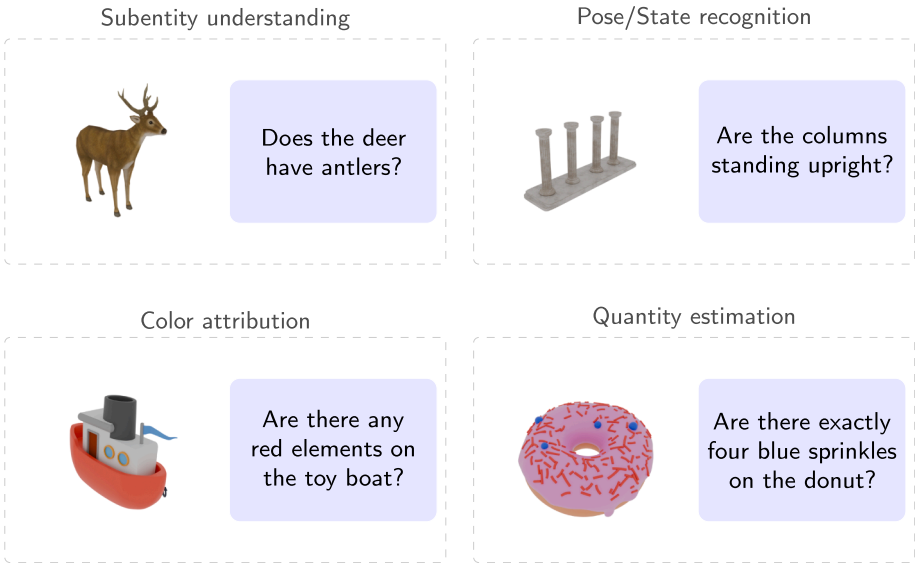

Fig. 5. Examples of each binary question type. Using these different question categories allows for a more fine grained assessment of the understanding capabilities of the model under evaluation. For example, many multimodal-LLMs struggle with counting questions (see Table 5).

This open-ended formulation aims to disentangle the model's semantic understanding from its instruction-following ability, offering a clearer view of the model's intrinsic comprehension skills. To assess the quality of generated answers, we employ an LLM as an automated judge, following trends established in recent MLLM evaluation studies [6,16,17], both in the 2D and 3D domains. Each predicted answer is compared against a human-written reference one, and the model's output is scored on a range between 0 and 100.

Captioning Task. To evaluate general understanding capabilities, we include an open-ended captioning task. Given a point cloud (or image) input, the model is prompted to produce a natural language description for the object. Similarly to the open-ended question-answering task, we employ an LLM as an automated judge to evaluate generated captions. The captioning task complements our binary and multiple-choice components by testing holistic understanding and linguistic fluency, offering insight into a model's capacity to extract and articulate structured semantic content from 3D input.

3.2. Shape categories

We categorize the 3D assets in our benchmark into four broad classes that reflect the diversity commonly found in digital content creation and virtual environments. These categories—*Creatures*, *Objects*, *Architecture*, and *Transport*-are derived from common taxonomies observed in large-scale asset repositories (e.g., SketchFab), promoting consistency with common workflows in computer graphics and similar settings. The *Creatures* category encompasses entities such as human characters, animals, and monster-like figures, all of which are central to computer graphics and video game development and require detailed structural and semantic understanding. The *Objects* category includes everyday items, such as electronics and household goods, which are relevant both in embodied AI contexts and in scene construction for interactive media. The *Architecture* category covers spatial structures like buildings, rooms, and staircases, commonly utilized in architectural design and engineering software, including tools such as Autodesk Inventor. Finally, the *Transport* category comprises various transportation modes, from bicycles to aircraft, frequently appearing in digital art and simulation-based applications.

This partitioning facilitates granular analysis of model performance with respect to geometric and semantic classes. In particular, we aim to identify systematic failure modes or category-specific limitations that may arise in 3D-LLMs. A more in-depth analysis of how performance across MLLMs may be affected by a specific shape category can be found in Appendix A.

3.3. Benchmark development stages

Due to the multimedia and multimodal nature of GLUE3D, its development required a multi-stage pipeline encompassing data preprocessing, point cloud generation, automatic question synthesis, and human annotation. A visual overview of the full development process is provided in Fig. 7.

STAGE I: Data Collection. As a first step, we select 128 textured polygon meshes from TurboSquid (https://www.turbosquid.com/), a large 3D asset-sharing platform. We deliberately choose TurboSquid over other sources (e.g., Sketchfab, CGTrader) because it is not included in the Objaverse dataset [19], which serves as the main source of training data for many 3D-LLMs. In addition, similar to the motivation in [19], TurboSquid offers a broader variety of styles and surface types than more traditional 3D datasets such as ModelNet [20] and ShapeNet [21].

Each of these is then processed: The mesh is normalized in a cube of size 1.0, while it is appropriately revolved so that the forward direction aligns with the positive y axis and the upward direction with the positive x axis. Incorrect texture and/or materials are also replaced with suitable ones, ensuring high-quality color information during the point cloud creation step. Finally, we convert each mesh into binary GLTF format, which makes it portable and (in the future) could allow the generation of different 3D representations, such as Gaussian splats [22], signed distance functions, and many others. All of the 128 meshes contain texture and material information, allowing for non-trivial surface signals.

STAGE II: Point Cloud and Image Generation. Using the Blender open-source software, we produce for each polygon mesh thirty random RGB-D views: Each mesh is normalized in a cube of size 1, and placed at the origin point. In particular, using the camera parameters, it is possible to transform the depth maps into global three-dimensional coordinates. These polygon meshes are then converted to point clouds using Blender: using the CYCLES rendering engine, we produce 30 random views of 512×512 RGB-D images (the depth channel uses half-precision floating

Fig. 6. Example of the multiple view renderings adopted for GLUE3D. The five views surround the objects at angles of 72°, with one always pointing towards the front of the object. The camera is at an angle of 30°, pointing downwards towards the origin, at a distance of 2.0 units (all surfaces are normalized to fit inside a unit cube).

point instead of a single byte). Using each view's camera parameters, we convert the depth channel to the world three-dimensional coordinates of the points, resulting in a collection of points $\hat{P}$, each with position coordinates and color coordinates (the latter containing discrete values between [0, 255]). To keep the size of the point cloud feasible, we downsample it using further-point-sampling to 50,000 points, which we denote as the set P. To compute the color of a point $p \in P$, we use the average color of all the points $\hat{p} \in \hat{P}$ whose closest point in P is p itself.

To generate the RGB images for 2D-LLMs evaluation, we select a representative view for each mesh, storing its respective camera parameters. From these, we render a 512×512 RGB image. Similarly to the point clouds included in our benchmark, the color channels for these images use discrete values between [0, 255]. Our multiple-view generation was done similarly to the single image setting, with the difference that five equidistant views are used instead of a human-provided one. Fig. 6 provides an example of a multiview rendering present in GLUE3D.

STAGE III: Binary Question Generation. To improve the development time of our benchmark, we decided to automatically generate plausible binary questions for each polygon mesh collected during STAGE I. To do so, we use the publicly available *Qwen3-30B-A3B* LLM [23]: This is a 30-billion, Mixture-of-Experts Model with reasoning capabilities. The generation procedure is as follows: for each polygon mesh, we manually produce a brief caption (e.g., *"a cat sitting"*, *"a car"*, …), and then prompt the language model to produce plausible binary questions.

STAGE IV: Question Labeling. Each binary question previously generated is then edited by a human annotator. Specifically, the annotator chooses whether or not a question is relevant to the input point cloud and, if so, annotates it with either a positive or negative answer. The annotator may also edit the question when they think it might be relevant. Building upon the annotated set of 1024 binary questions, we generate a set of 256 multiple-choice questions using the Qwen3-30B-A3B language model. For each 3D shape in the benchmark, the model is prompted to generate two multiple-choice questions, using the human-annotated labels. Finally, each multiple-choice question-answer pair is manually validated and reformulated into an open-ended format, resulting in our open-ended task.

4. Experimental setup

4.1. Task design

Question answering procedure. All MLLMs evaluated in our benchmark are instruction-tuned and designed to operate within a conversational, user-assistant framework typical of modern chatbot interfaces. Accordingly, for each question in our benchmark, we first provide the input point cloud (or image) normalized to match the specific input requirements of the target MLLM, then we embed the question as part of a user message. Binary question responses are obtained by substituting each question into the prompt template shown in Table 8, and selecting the token with the highest predicted probability between `Yes` and `No`. For multiple-choice questions, the procedure is analogous: each prompt is evaluated by computing the next-token probabilities over the four candidate choices: `A`, `B`, `C`, and `D`, and selecting the option with the highest score. For the captioning and open-ended tasks, sampling is performed by iteratively selecting the next most probable token.

Table 2
Prompt-adherence percentage for GLUE3D binary and multiple-choice tasks. While most models show robust performance on binary questions without requiring explicit generation constraints, their performance deteriorates for multiple-choice queries. This is especially true for current 3D-LLMs, suggesting a struggle with instruction-following tasks that involve higher structural or semantic complexity.

Model	Q&A Task	
	Binary	Multiple-choice
3D-LLMs		
PointLLM v1.2 7B	99.90 %	5.86 %
PointLLM v1.2 13B	99.71 %	1.95 %
MiniGPT-3D	0.00 %	0.00 %
ShapeLLM 7B	100.00 %	0.39 %
2D-LLMs		
Qwen2.5-VL 7B instruct	100.00 %	100.00 %
Phi 3.5 vision-instruct	100.00 %	100.00 %
LlaVa 1.5 7B	100.00 %	100.00 %
Multiview LLMs		
Qwen2.5-VL 7B instruct	100.00 %	100.00 %
Phi 3.5 vision-instruct	100.00 %	100.00 %
LlaVa 3D	100.00 %	98.05 %

Prompt-adherence for binary and multiple-choice questions. The advantages of using binary and multiple-choice question is the lack of sophisticated evaluation procedure, which may hinder the benchmark applicability and longevity. However, to properly work, the model under scrutiny must have the necessary instruction-following capabilities to follow the requested format. Unfortunately, as can be observed in Table 2, when generating without constraints, 3D-LLMs seem to struggle in following the required output format. This is especially true for the multiple-choice task. Furthermore, since both ShapeLLM, PointLLM, and LlaVa 1.5 share the same textual backbone, the observation that only the first two models exhibit difficulties in following instructions suggests that either the 3D multimodal alignment is suboptimal or that catastrophic forgetting may have occurred during the multimodal fine-tuning phase. Either way, as can be seen from the 2D-LLMs performance, we should expect a sufficiently instruction-tuned model to easily be able to adapt to our instruction prompts.

Textual baseline. To better contextualize the performance of vision- and geometry-based MLLMs, we include a textual-only baseline in our evaluation. This baseline serves to quantify how much of the task can be completed using language priors alone, without access to visual or geometric input. To construct this baseline, we provide to a large language model only a brief textual caption associated with the 3D object (e.g., *"a cat sitting"*, *"a red sports car"*), and prompt it with the same binary or

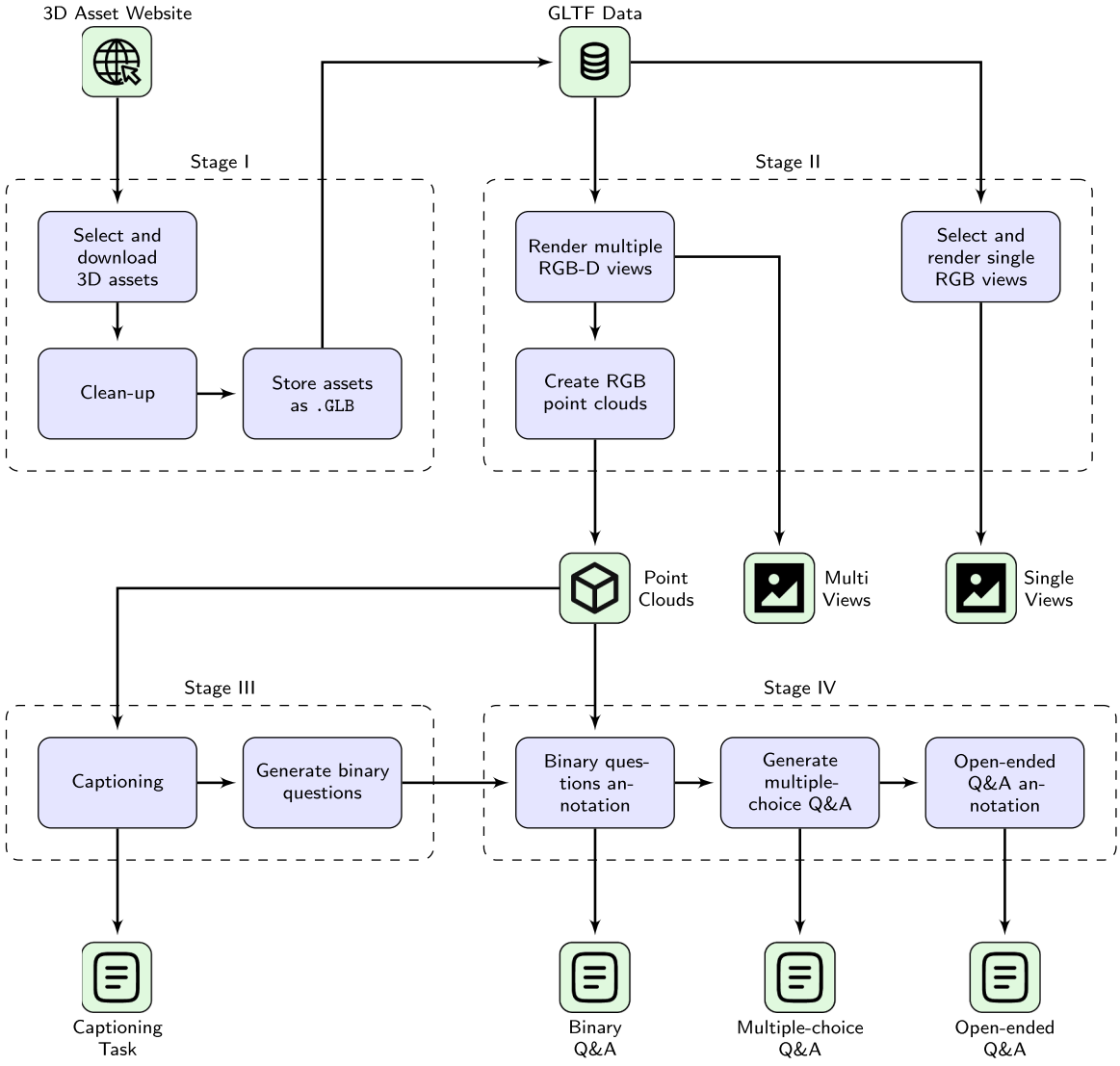

Fig. 7. Construction pipeline of GLUE3D. The process unfolds in four stages. Stage I: 3-D assets are selected and downloaded from a public repository, cleaned, and saved in GLB format. Stage II: Each asset is converted to two complementary representations: (i) multi-view RGB-D renderings that are fused into RGB point clouds, and (ii) a curated single RGB view for image-based evaluation. Stage III: Caption prompts and a pool of candidate binary questions are automatically generated for every asset. Stage IV: Experts annotate the binary questions and author multiple-choice items, completing the captioning, binary, and multiple-choice subsets that make up the final benchmark.

multiple-choice question used in the full benchmark. No visual or point cloud data is supplied. The caption is manually written by human annotators. Because the caption encodes high-level semantic information—but omits spatial or appearance-specific details—this setup tests whether models may infer plausible answers based solely on the object class.

4.2. Models

Given the multimodal nature of our benchmark, we evaluate a diverse set of models spanning various input modalities. In particular, it is useful to differentiate between *3D-LLMs*, *2D-LLMs*, and traditional unimodal *LLMs*. Table 3 gives an overview of the size and modalities for each tested model.

3D models. For 3D-LLMs, we use the current *state-of-the-art* models in object-oriented question answering: These include PointLLM v1.2 7B, PointLLM v1.2 13B [16], ShapeLLM 7B [17], and minigptd [24].

The PointLLM models use Vicuna v1.1 [14] (7B and 13B respectively) as their textual backbone, while the point cloud encoder consists of Point-BERT [25]. It adopts a two-stage training approach; first, learning projection matrices between point-BERT and Vicuna, then fine-tuning everything together. Similarly to PointLLM, ShapeLLM [17] also adopts Vicuna v1.1 as their language backbone; however, rather than using Point-BERT as modality encoder, they opt for a variation of a more recent (and seemingly improved) point cloud encoder, i.e., ReCon [26]. MiniGPT-3D adopts a more parameter-conscious approach and works by updating an existing 2D and 3D priors (BLIP-2 and Point-BERT, respectively) and aligning them with the Phi-2 [27] language backbone. This results in a surprisingly powerful yet efficient 3D-LLM. Finally, to bridge 3D and 2D models, we include LLaVA-3D [28], a multimodal language model that uses RGBD image sequences as its input. LLaVA-3D leverages LlaVa 1.5 7B [29] 2D priors to build its three-dimensional scene understanding capabilities. This is achieved by injecting depth information using 3D positional encodings, thus providing more information to

Binary Task

"Answer only with "Yes" or "No": {question}?"

Multiple-choice Task

"Only answer with either A,B,C,D: {question}?
A) {choice_a}
B) {choice_b}
C) {choice_c}
D) {choice_d}"

Open-ended Task

"Answer the following question: {question}?"

Captioning Task

"What is this?"

Fig. 8. GLUE3D Task Templates. Prompts used for our benchmark tasks: At test time, blue text is substituted with the appropriate content. Each prompt is wrapped in a user message and dispatched to the MLLMs under scrutiny along with its respective input point cloud (or image).

the language backbone. Empirical results indicate that LLaVA-3D offers greater 3D understanding compared to many prior works [30–32].

2D models. For our 2D-LLMs, we provide baseline data from the following models: LlaVa 1.5 7B [29], phivision [33], and qwenvl [34], as they present strong baselines for current medium-sized MLLMs capabilities.

LlaVa 1.5 7B was among the first image-conditioned Large Language Models, and utilizes a relatively simple architecture. It learns a set of vision tokens (obtained from the projections of a frozen vision encoder) to prepend to the main textual body. Learning data consists of 158K instruction-following samples. As the final training stage they fine-tune end-to-end both the projection matrices and the LLM backbone. Phi 3.5 vision-instruct is part of the Phi-3.5 family of LLMs, developed by Microsoft. It utilizes CLIP as its visual encoder and Phi-3.5-mini-instruct as its textual backbone. The model is pre-trained on a multimodal dataset of 0.5 trillion tokens and fine-tuned on a dataset of 33 billion tokens. The Qwen 2.5 VL family is the latest among the above models; it uses several sophisticated techniques to handle a variety of media, from images to videos. Fittingly, it utilizes Qwen 2.5 LLMs and a custom ViT encoder architecture. These are then pre-trained on large quantities of multimodal data (~ 4 trillion tokens) and fine-tuned on two million samples of complex single-turn and multi-turn instruction-following data.

Textual only models. For textual only models, we use Phi-3.5-mini-instruct [33] and Llama-3-instruct [35]. These are instruction-tuned variants of their respective pretrained language models. For this benchmark, these models are used to provide the expected answer using only minimal textual information. Furthermore, we include both Vicuna v1.1 7B and Vicuna v1.1 13B: Since they act as backbones for both PointLLM and ShapeLLM, having an idea of their general understanding concerning GLUE3D tasks may be insightful. Finally, we include LLaMA-Mesh [36], a language model fine-tuned to unify textual and mesh representations, as a more spatially grounded textual baseline. Its built-in spatial understanding capabilities make it suitable for providing more general and spatially informed answers.

Table 3
Overview of the language models employed in our analysis. As most existing 3D-LLMs fall within the small to medium parameter range (approximately 3B-13B), we also selected 2D-LLMs and text-only LLMs of comparable size to ensure a fair evaluation.

Model	Parameters	Modalities
PointLLM v1.2 7B	7 Billion	Point Cloud-Text
PointLLM v1.2 13B	13 Billion	Point Cloud-Text
MiniGPT-3D	2.7 Billion	Point Cloud-Text
ShapeLLM 7B	7 Billion	Point Cloud-Text
ShapeLLM 13B	13 Billion	Point Cloud-Text
LlaVa 3D	7 Billion	RGBD Images-Text
Qwen2.5-VL 7B instruct	8 Billion	Image-Text
Phi 3.5 vision-instruct	4.2 Billion	Image-Text
LlaVa 1.5 7B	7 Billion	Image-Text
Llama 3 8B instruct	8 Billion	Text-only
Phi 3.5 mini-instruct	2.7 Billion	Text-only
Vicuna v1.1 7B	7 Billion	Text-only
Vicuna v1.1 13B	13 Billion	Text-only
LlaMA-Mesh	8 Billion	Text-only

4.3. Metrics

To assess model performance across the different task types in our benchmark, we employ both classification and open-ended evaluation metrics: For the binary and multiple-choice tasks, we report commonly used classification metrics: *Accuracy*, *Precision*, *Recall*, and *F1 score*. In the multiple-choice scenario, Precision and Recall are computed as the average of multiple—one per choice—binary classification tasks. In contrast, the open-ended question-answering and the captioning task are inherently generative, making traditional classification metrics unsuitable. To address this, we adopt an *LLM-as-a-judge* framework, following prior work [6,16,17], wherein a language model evaluates the correctness of the generated answers. Prompts used by our evaluation judge are reported in Appendix B. Consistent with prior work [16], we include a description of the question-answer evaluation task, together with representative scoring examples. For the sake of reproducibility, the language model of our choice is *Qwen3-30B-A3B* [23], a powerful open-source language model with impressive generation and understanding capabilities. Additionally, to provide a broader comparison of caption quality, we compute sentence-level similarity scores using established embedding-based methods, namely *Sentence-BERT* [37] and *SimCSE* [38], and more traditional n-gram based metrics such as BLEU [39], METEOR [40], and ROUGE [41].

Table 4
Classification results for the binary question-answering task. This task exhibits the largest performance gap between 3D-LLMs and 2D-LLMs. In terms of accuracy, most 3D-LLMs fail to surpass even the random baseline of 0.5 or the textual-only baseline, suggesting significant limitations in their current perception or instruction-following capabilities.

Binary Questions Model	Accuracy	Precision	Recall	F1-score
3D-LLMs				
PointLLM v1.2 7B	0.485	0.488	0.866	0.624
PointLLM v1.2 13B	0.490	0.491	0.909	0.638
ShapeLLM 7B	0.494	0.494	1.000	0.661
MiniGPT-3D	0.548	0.536	0.632	0.580
2D-LLMs				
Qwen2.5-VL 7B instruct	0.796	0.888	0.672	0.765
Phi 3.5 vision-instruct	0.728	0.707	0.765	0.735
LlaVa 1.5 7B	0.682	0.630	0.864	0.728
Multiview LLMs				
Qwen2.5-VL 7B instruct	0.798	0.914	0.652	0.761
Phi 3.5 vision-instruct	0.716	0.683	0.792	0.734
LlaVa 3D	0.662	0.627	0.783	0.696
Textual baselines				
Phi 3.5 mini-instruct	0.591	0.623	0.435	0.512
Llama 3 8B instruct	0.608	0.610	0.575	0.592
Vicuna v1.1 7B	0.549	0.553	0.451	0.497
Vicuna v1.1 13B	0.559	0.539	0.729	0.620
LlaMA-Mesh	0.579	0.569	0.611	0.589

5. Results and analysis

5.1. Quantitative comparison

We report quantitative performance of all evaluated models across the four core tasks of our benchmark: binary question-answering, multiple-choice selection, open-ended question-answering, and caption generation. Table 4 introduces the results for the binary classification task, including accuracy, precision, recall, and F1 score for each model. These metrics offer insight into the models' ability to correctly identify fine-grained properties in point clouds and images. This is further analyzed in Tables 5, 6, which provide model results across various questions and 3D geometry types. Tables 7, 8 show the performance on the multiple-choice task, and its open-ended variant. Finally, Tables 9, 10 report the results of the captioning task, where generated captions are compared against ground-truth descriptions. Together, these tables provide a comprehensive view of the strengths and weaknesses of each model across different modalities and task formats.

5.2. Discussion

3D vs 2D. As can be seen in Tables 4 to 10, 2D-LLMs systematically outperform 3D-LLMs in all settings. It is difficult to assess the causes of this imbalance; it might be related to the larger volume of low- and high-quality training data, or maybe less effective point cloud encoding architectures with respect to the image ones. Similar textual backbones are adopted by both 2D-LLMs and 3D-LLMs, thus it is unlikely that to be the cause of this performance deficit.

Multiview setting. Interestingly, the additional views—while providing a slight improvement with respect to the single view setting—do not particularly improve performance of current 2D-LLMs models. This may suggest that, for most cases, the supplementary information contained in the different views is not required to produce satisfactory results on GLUE3D. In particular, Qwen2.5-VL 7B instruct is often able to properly exploit the additional information provided by the multiple views, as can be clearly seen in the multiple-choice and open-ended tasks (Tables 7 and 8). On the contrary, Phi 3.5 vision-instruct performance often worsens when adopting multiple views, suggesting that the additional information is actually detrimental and may be causing the model to misunderstand the input objects.

Table 5
Accuracy scores across various types of binary probing questions. While question type does not appear to affect 3D-LLMs significantly, 2D-LLMs show greater difficulty with quantification and counting questions.

Binary Questions Model	Color	Counting	Pose	Subentity
3D-LLMs				
PointLLM v1.2 7B	0.492	0.494	0.461	0.492
PointLLM v1.2 13B	0.492	0.498	0.457	0.511
ShapeLLM 7B	0.492	0.494	0.486	0.504
MiniGPT-3D	0.619	0.530	0.482	0.556
2D-LLMs				
Qwen2.5-VL 7B instruct	0.800	0.723	0.788	0.868
Phi 3.5 vision-instruct	0.758	0.632	0.747	0.771
LlaVa 1.5 7B	0.708	0.538	0.735	0.744
Multiview LLMs				
Qwen2.5-VL 7B instruct	0.835	0.759	0.767	0.827
Phi 3.5 vision-instruct	0.792	0.617	0.686	0.763
LlaVa 3D	0.723	0.534	0.645	0.741
Textual baselines				
Phi 3.5 mini-instruct	0.542	0.613	0.604	0.605
Vicuna v1.1 7B	0.512	0.565	0.559	0.560
Vicuna v1.1 13B	0.565	0.573	0.539	0.556
Llama 3 8B instruct	0.550	0.589	0.653	0.643
LlaMA-Mesh	0.523	0.561	0.624	0.609

Table 6
Accuracy scores for both the binary and multiple-choice Q&A tasks across various classes of surfaces; Architecture (Ar), Creatures (Cr), Objects (Ob), and Transport (Tr). For the models under evaluation, surface type does not appear to significantly influence performance.

	Binary				Multiple-Choice			
Model	Ar	Cr	Ob	Tr	Ar	Cr	Ob	Tr
3D-LLMs								
PointLLM v1.2 7B	0.52	0.43	0.48	0.50	0.33	0.23	0.27	0.41
PointLLM v1.2 13B	0.48	0.49	0.51	0.47	0.38	0.30	0.34	0.38
ShapeLLM 7B	0.50	0.49	0.50	0.48	0.30	0.28	0.27	0.39
MiniGPT-3D	0.56	0.54	0.56	0.54	0.25	0.30	0.25	0.31
2D-LLMs								
Qwen2.5-VL 7B instruct	0.80	0.81	0.81	0.76	0.72	0.81	0.73	0.73
Phi 3.5 vision-instruct	0.74	0.78	0.71	0.67	0.62	0.69	0.66	0.62
LlaVa 1.5 7B	0.72	0.70	0.67	0.64	0.56	0.58	0.55	0.55
Multiview LLMs								
Qwen2.5-VL 7B instruct	0.79	0.84	0.81	0.75	0.81	0.77	0.73	0.81
Phi 3.5 vision-instruct	0.72	0.75	0.69	0.71	0.64	0.64	0.67	0.75
LlaVa 3D	0.66	0.71	0.65	0.64	0.48	0.50	0.48	0.50
Textual baselines								
Phi 3.5 mini-instruct	0.53	0.59	0.63	0.61	0.42	0.44	0.50	0.52
Vicuna v1.1 7B	0.55	0.57	0.52	0.56	0.42	0.34	0.45	0.38
Vicuna v1.1 13B	0.56	0.54	0.56	0.58	0.31	0.38	0.38	0.44
Llama 3 8B instruct	0.57	0.59	0.63	0.64	0.50	0.50	0.62	0.56
LlaMA-Mesh	0.52	0.60	0.59	0.61	0.47	0.53	0.62	0.55

3D vs text. Interestingly enough, 3D-LLMs seem to struggle even with respect to textual models. This suggests that either they are not able to categorize the input geometries correctly, or they do not properly harness their language backbone to provide the most probable answer. This can be seen both for the binary and multiple-choice tasks Tables 4, 7 respectively. On the contrary, 2D-LLMs do not suffer the same problem.

Main limitations of the available models. Many MLLMs seem to struggle in instruction-following tasks, as can be observed by their poor results in the binary and multiple-choice tasks; Indeed, some models tend to systematically choose to answer positively for binary questions, and with

Table 7
Classification results for the multiple-choice question-answering task. Similar to the binary question-answering task, 3D-LLMs also exhibit difficulty on multiple-choice questions, with their performance consistently falling short of both 2D-LLMs and textual-only baselines.

Multiple-Choice Questions Model	Accuracy	Precision	Recall	F1-score
3D-LLMs				
PointLLM v1.2 7B	0.309	0.407	0.309	0.261
PointLLM v1.2 13B	0.348	0.436	0.348	0.320
MiniGPT-3D	0.277	0.304	0.277	0.202
ShapeLLM 7B	0.309	0.600	0.309	0.220
2D-LLMs				
Qwen2.5-VL 7B instruct	0.750	0.760	0.750	0.752
Phi 3.5 vision-instruct	0.648	0.669	0.648	0.649
LlaVa 1.5 7B	0.559	0.633	0.559	0.555
Multiview LLMs				
Qwen2.5-VL 7B instruct	0.781	0.786	0.781	0.782
Phi 3.5 vision-instruct	0.676	0.689	0.676	0.676
LlaVa 3D	0.492	0.621	0.492	0.483
Textual baselines				
Phi 3.5 mini-instruct	0.469	0.470	0.469	0.469
Vicuna v1.1 13B	0.375	0.427	0.375	0.369
Vicuna v1.1 7B	0.398	0.494	0.398	0.369
Llama 3 8B instruct	0.547	0.575	0.547	0.545
LlaMA-Mesh	0.543	0.577	0.543	0.538

Table 8
Judge scores for the open-ended question-answering task. We report both the overall average score of each model and the per-category results across the four surface categories: Architecture (Ar), Creatures (Cr), Objects (Ob), and Transport (Tr). Similar to other tasks in our benchmark, 2D-LLMs consistently achieve better performance. This seems to corroborate our hypothesis of a gap between current 2D-LLMs *state-of-the-art* models and their three-dimensional counterparts.

Open-Ended Questions Model	Ar	Cr	Ob	Tr	Average
3D-LLMs					
PointLLM v1.2 7B	52.97	68.44	63.05	64.53	62.25
PointLLM v1.2 13B	52.42	67.11	68.05	61.25	62.21
ShapeLLM 7B	53.67	67.73	61.33	63.28	61.50
MiniGPT-3D	58.91	70.23	65.94	64.06	64.79
2D-LLMs					
Qwen2.5-VL 7B instruct	69.84	79.30	74.53	75.55	74.80
Phi 3.5 vision-instruct	68.72	77.34	72.50	75.00	73.39
LlaVa 1.5 7B	67.39	84.06	75.08	72.66	74.80
Multiview 2D-LLMs					
Qwen2.5-VL 7B instruct	77.66	82.27	77.19	75.86	78.24
Phi 3.5 vision-instruct	66.84	74.22	73.05	74.22	72.08
LlaVa 3D	69.03	79.45	72.50	73.12	73.53

Table 9
Results for the GLUE3D captioning task. Interestingly, 3D-LLMs exhibit a smaller performance gap on the captioning task compared to other evaluation settings. This may be attributed to the fact that a significant portion of the training data [42] for the PointLLM, ShapeLLM, and MiniGPT-3D models consists of relatively simple captioning instructions.

Captioning Task Model	Qwen3-as-a-Judge	S-BERT	SimCSE
3D-LLMs			
PointLLM v1.2 7B	56.680	50.612	46.860
PointLLM v1.2 13B	53.750	49.267	47.158
ShapeLLM 7B	38.047	36.012	30.524
MiniGPT-3D	55.859	51.877	46.249
2D-LLMs			
Qwen2.5-VL 7B instruct	68.867	65.558	65.277
Phi 3.5 vision-instruct	67.070	58.629	60.112
LlaVa 1.5 7B	64.688	59.686	60.013
Multiview LLMs			
Qwen2.5-VL 7B instruct	69.531	64.598	63.821
Phi 3.5 vision-instruct	66.758	58.993	60.172
LlaVa 3D	57.109	40.346	44.349

Table 10
Additional results for the GLUE3D captioning task, evaluated using traditional n-gram-based NLP metrics. We report BLEU (B), ROUGE-L (R-L), and METEOR (M) scores for all evaluated models, grouped by modality.

Captioning Task Model	B-1	B-2	B-3	R-L	M
3D-LLMs					
PointLLM v1.2 7B	6.935	2.162	1.014	10.569	14.619
PointLLM v1.2 13B	6.665	2.196	1.071	10.477	14.235
ShapeLLM 7B	7.395	2.286	1.344	9.982	8.835
MiniGPT-3D	12.091	5.837	3.216	16.091	17.830
2D-LLMs					
Qwen2.5-VL 7B instruct	10.766	4.674	2.455	16.139	20.981
Phi 3.5 vision-instruct	9.442	4.087	2.573	16.134	11.490
LlaVa 1.5 7B	11.057	4.919	2.444	17.262	18.932
Multiview LLMs					
Qwen2.5-VL 7B instruct	9.419	4.308	2.163	14.533	20.106
Phi 3.5 vision-instruct	10.627	4.533	2.687	16.795	12.516
LlaVa 3D	0.051	0.015	0.012	0.443	0.643

choice `A` for multiple-choice. We are not the first to observe this result, as a similar occurrence was also reported in [3]. This is consistent with what we observed in Table 2; several 3D-LLMs have difficulties in precisely following the user-provided instructions, more so than current state-of-the-art 2D-LLMs. Thus, an important direction for future 3D-LLMs should be the development of models able to precisely follow instructions.

Main limitations of GLUE3D and future directions. While GLUE3D offers a structured and interpretable evaluation framework for 3D-LLMs, it also presents several limitations that we aim to address in future work. First, the current set of tasks-particularly binary and multiple-choice formats-relies heavily on instruction-tuned capabilities, which may bias evaluation toward models optimized for chat-based interactions. Future iterations could incorporate tasks that are less instruction-sensitive, such as contrastive or retrieval-based setups. Second, the benchmark currently includes a limited number of surfaces—only 128; expanding the dataset to cover a broader range of shapes, textures, and structural complexity would improve diversity and robustness. Finally, most of the current tasks focus on perceptual understanding. Extending the benchmark to include higher-level reasoning tasks would enable deeper evaluation of 3D-LLMs beyond object-level understanding. We view these directions as key steps toward a more comprehensive and generalizable multimodal benchmark on point clouds.

A further compelling future direction would be to leverage GLUE3D resources (3D data and textual annotations), to develop broader multimodal benchmarks, making possible the evaluation of different types of multimodal models and architectures than solely 2D/3D LLMs. For example, the development of methodologies for evaluating multimodal self-supervised approaches, as well as surface (or image) generative models, could be a worthwhile endeavor.

6. Conclusion

In this work, we introduced GLUE3D, a novel benchmark designed to evaluate the capabilities of multimodal large language models in understanding 3D point cloud data. Our benchmark provides a diverse set of evaluation tasks-including binary questions, multiple-choice questions, and open-ended questions-enabling fine-grained and holistic assessment across different perception modalities. Furthermore, by aligning 3D data with their 2D projections, GLUE3D offers a unified framework for probing both 2D and 3D language models, effectively quantifying their difference in performance.

Through this benchmark, we aim to shed light on the limitations of current 3D-LLMs in spatial, geometric, and semantic understanding. By presenting every asset as both a colour-rich point cloud and its matched

RGB rendering, GLUE3D compels multimodal LLMs to reconcile complementary geometric and photometric cues within a single generative framework, effectively turning evaluation itself into information fusion tasks. This dual-view design, together with tasks that demand spatial reasoning, attribute binding, and free-form description, offers a systematic lens on how well current models integrate heterogeneous sources rather than relying on text-only priors. In doing so, the benchmark supplies not just fresh data but a principled yardstick for pushing scalable, geometry-aware fusion methods grounded in generative AI.

CRediT authorship contribution statement

Giorgio Mariani: Writing - original draft, Methodology, Investigation, Data curation, Conceptualization; **Alessandro Raganato:** Writing - review & editing, Supervision, Investigation; **Simone Melzi:** Writing - review & editing, Supervision; **Gabriella Pasi:** Writing - review & editing, Supervision, Funding acquisition.

Data availability

The evaluation scripts used in this work are available at https://github.com/giorgio-mariani/GLUE3D. All necessary data is automatically retrieved by these scripts. The full dataset, including GLTF files with mesh and texture information, can also be accessed directly at https://huggingface.co/datasets/giorgio-mariani-1/GLUE3D.

Declaration of generative AI in the writing process

During the preparation of this work, the authors used ChatGPT in order to rephrase and improve the content of this paper. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding and Acknowledgment

This work was partially supported by the European Union - Next Generation EU within the project PNRR M4C2, Investment 1.3 DD. 341 - 15 March 2022 - FAIR - Future Artificial Intelligence Research - Spoke 4 - PE00000013 - D53C22002380006.

Appendix A. Category analysis

In the main paper, we report evaluation metrics across four semantic categories (Architecture, Creatures, Objects, and Transport) to gain a deeper understanding of the models' understanding capabilities over different types of surfaces. However, the variations in performance observed across categories may also arise from external factors, such as differences in geometric complexity or data distribution, rather than from inherent model limitations. To better interpret these category-wise discrepancies, we therefore conduct a simple comparative analysis using the PointLLM Objaverse captioning task [16]

We manually annotated each of the 200 point clouds in the Objaverse captioning task according to our proposed categories. The resulting distribution is illustrated in Fig. A.1. Following the same evaluation procedure described in Section 4.3, we employ *Qwen3-30B-A3B* as an automated judging model to assess the quality of model-generated captions. The comparative results for both GLUE3D and Objaverse are presented in Table A.1. From this analysis, we observe two main trends:

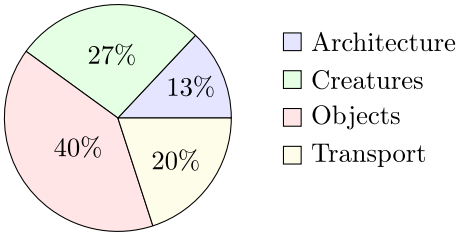

Fig. A.1. Distribution of point cloud categories in Objaverse. Differently from GLUE3D, the object categories in Objaverse are not evenly distributed. Indeed, of the 200 point clouds in the benchmark, 26 are architectural structures, 54 are creatures and characters, 80 are everyday objects, and 40 are real and fictional vehicles. These categories were manually annotated by us, and can be found at the following URL: https://huggingface.co/datasets/giorgio-mariani-1/GLUE3D/resolve/main/annotations/categories/categories_Objaverse200.jsonl.

Table A.1
Judge scores on both GLUE3D and the Objaverse captioning tasks. It is interesting to observe that minima (underlined) and maxima (bold) scores seem to be aligned across 3D-LLMs, suggesting a possible bias in the 3D training datasets, as PointLLM, ShapeLLM, and MiniGPT-3D share portions of their training data.

Objaverse Captioning Task Model	Ar	Cr	Ob	Tr
PointLLM v1.2 7B	51.77	52.20	53.02	**54.95**
PointLLM v1.2 13B	51.92	54.94	52.24	**56.50**
ShapeLLM 7B	48.58	46.39	48.19	**50.62**
MiniGPT-3D	54.15	54.37	53.40	**58.98**
GLUE3D Captioning Task				
PointLLM v1.2 7B	49.84	55.31	58.75	**62.81**
PointLLM v1.2 13B	50.31	51.88	56.25	**56.56**
ShapeLLM 7B	32.50	33.12	38.12	**48.44**
MiniGPT-3D	49.69	57.50	55.78	**60.47**
LlaVa 3D	52.50	56.25	59.06	**60.62**
Qwen2.5-VL 7B instruct	66.88	**70.62**	68.28	69.69
Phi 3.5 vision-instruct	63.44	63.12	**71.41**	70.31
LlaVa 1.5 7B	63.91	**66.88**	64.84	63.12

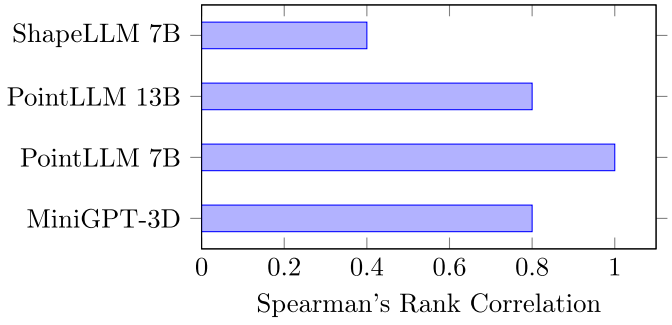

Fig. A.2. Spearman's rank correlation between mean category scores of various 3D-LLMs. This figure presents the Spearman correlation between the category-level judge scores obtained on Objaverse and those on GLUE3D. As also shown in Table A.1, the performance of most 3D-LLMs exhibits a strong positive correlation across datasets, suggesting a shared underlying factor influencing category-wise performance.

- 3D-LLMs exhibit correlated performance across both benchmarks and models, suggesting consistent category-specific challenges.
- 2D-LLMs (which can be evaluated only on GLUE3D) do not show such correlations, indicating that the underlying cause is likely specific to 3D models-potentially due to factors such as incorrect or insufficient training data, suboptimal point-cloud encoders, or limitations in multimodal alignment (Fig. A.2).

```
"Evaluate a model-generated caption against a human-provided caption (ground truth). The former is the
output from a captioning system; the latter is written by a human annotator.

Consider the two captions and provide a score representing how coherent the two sentences are with
each other. Your response should consist of a single confidence score ranging from 0 to 100.

Below are several examples of question-answer pairs along with their corresponding confidence scores:

question1: How many oranges will there be if 1/3 of them are removed?
answer from model: There will be 6 left.
answer from label: As there are 9 oranges in total, there will be 6 oranges left if 1/3 of them are
removed.
confidence score: 100

question2: What is this object?
answer from model: This is a bathtub
answer from label: This is a dirty bathtub.
confidence score: 80

question3: What is this object?
answer from model: This is a bottle of water
answer from label: This is a bottle of oil
confidence score: 50

question4: What does the boy have in his right hand?
answer from model: He is holding a white cup in his right hand.
answer from label: He is holding a sword in his right hand.
confidence score: 0

Next, I will give you the following inputs:
question: What is this object?
answer from model: {model-output}
answer from label: {ground-truth}

Please remember: your output should contain **only** the confidence score, no words or punctuation,
just the number."
```

Fig. B.1. LLM-as-a-judge evaluation template. We adopt the above prompt template for scoring captioning results in GLUE3D.

Appendix B. LLM-as-a-judge prompts

As described in Section 4.3, we adopt an LLM-as-a-Judge approach for evaluation of our open-ended and captioning tasks. The prompts used by our judge can be seen in Figs. B.1, B.2. They are based on prompts adopted in previous work [16] (Fig. B.2).

```
“Evaluate a model-generated answer against a human-provided answer (ground truth).
The former is the output from a Visual Q&A system; the latter is written by a human annotator.

Consider the two answers and provide a score representing how consistent is the model-generated answer
with the human-provided one.
Your response should consist of a single confidence score ranging from 0 to 100.
Focus on the information content provided in model-generated answer.  Is it consistent with the ground
truth?  Is it not?

The score should **not** be influenced by:
- The syntax of the answer; "There are three sticks." is equivalent to "3 sticks".
- The presence of additional information in model generated answer should not negatively influence the
score.
- The confidence of the answer should not negatively influence the score:  "This appears to be an apple"
should be score the same as "This is an apple."
- Grammar errors should not influence the confidence score.

--

Below are several examples of question-answer pairs along with their corresponding confidence scores:
question:  How many oranges will there be if 1/3 of them are removed?
answer from model:  There will be 6 left.
answer from label:  As there are 9 oranges in total, there will be 6 oranges left if 1/3 of them are
removed.
confidence score:  100

question:  How many wheels doe the car have?
answer from model:  3
answer from label:  The car has three wheels.
confidence score:  100

question:  What is the content of the bottle?
answer from model:  This is a bottle of water.
answer from label:  This is a bottle of oil.
confidence score:  0

question:  What does the boy have in his right hand?
answer from model:  He is holding a white cup in his right hand.
answer from label:  The boy is holding a white glass in his right hand.
confidence score:  50

--

Next, estimate the confidence score for the following answer pair:
question:  {question}?
answer from model:  {model-output}
answer from label:  {ground-truth}

Please remember, your output should contain **only** the confidence score, no words or punctuation,
just the number.”
```

Fig. B.2. LLM-as-a-judge evaluation template. We adopt the above prompt template for scoring answers in GLUE3D.

References

[1] S. Leng, H. Zhang, G. Chen, X. Li, S. Lu, C. Miao, L. Bing, Mitigating object hallucinations in large vision-language models through visual contrastive decoding, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR, 2023, pp. 13872–13882.

[2] S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. Lecun, S. Xie, Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9568–9578.

[3] Y. Li, Y. Du, K. Zhou, J. Wang, W.X. Zhao, J.-R. Wen, Evaluating object hallucination in large vision-language models, in: The 2023 Conference on Empirical Methods in Natural Language Processing, 2023.

[4] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, Z. Qiu, W. Lin, J. Yang, X. Zheng, K. Li, X. Sun, R. Ji, MME: a comprehensive evaluation benchmark for multimodal large language models,, 2023. abs/2306.13394

[5] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, et al., MMBench: is your multi-modal model an all-around player?, in: European Conference on Computer Vision, 2024, pp. 216–233.

[6] W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, L. Wang, MM-Vet: evaluating large multimodal models for integrated capabilities, in: Proceedings of the 41st International Conference on Machine Learning, 2024, pp. 57730–57754.

[7] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. Mccandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, Inc, 2020.

[8] Z. Bai, P. Wang, T. Xiao, T. He, Z. Han, Z. Zhang, M.Z. Shou, Hallucination of multimodal large language models: a survey, 2024. abs/2404.18930

[9] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, D. Parikh, Making the V in VQA matter: elevating the role of image understanding in visual question answering, Int. J. Comput. Vis. 127 (2016) 398–414.

[10] T.-Y. Lin, M. Maire, S.J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: Common objects in context, 2014. European Conference on Computer Vision.

[11] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, P. Anderson, NoCaps: novel object captioning at scale, International Conference on Computer Vision, 2019.

[12] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, M. Rohrbach, Towards VQA models that can read, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR, 2019, pp. 8309–8318.

[13] O. Sidorov, R. Hu, M. Rohrbach, A. Singh, TextCaps: a dataset for image captioning with reading comprehension, in: European Conference on Computer Vision, 2020, pp. 742–758.

[14] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al., Judging LLM-as-a-judge with MT-bench and chatbot arena, Adv. Neural Inf. Process. Syst. 36 (2023) 46595–46623.

[15] J. Yang, X. Chen, N. Madaan, M. Iyengar, S. Qian, D.F. Fouhey, J. Chai, 3D-GRAND: a million-scale dataset for 3D-LLMs with better grounding and less hallucination, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 29501–29512.

[16] R. Xu, X. Wang, T. Wang, Y. Chen, J. Pang, D. Lin, PointLLM: empowering large language models to understand point clouds, in: European Conference on Computer Vision, 2023.

[17] Z. Qi, R. Dong, S. Zhang, H. Geng, C. Han, Z. Ge, L. Yi, K. Ma, ShapeLLM: universal 3D object understanding for embodied interaction, Springer, 2024. European Conference on Computer Vision.

[18] L. Chen, J. Li, X. Dong, P. Zhang, Y. Zang, Z. Chen, H. Duan, J. Wang, Y. Qiao, D. Lin, et al., Are we on the right way for evaluating large vision-language models?, Adv. Neural Inf. Process. Syst. 37 (2024) 27056–27087.

[19] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. Vanderbilt, L. Schmidt, K. Ehsani, A. Kembhavi, A. Farhadi, Objaverse: a universe of annotated 3D objects, 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[20] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, J. Xiao, 3D ShapeNets: a deep representation for volumetric shapes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[21] A.X. Chang, T.A. Funkhouser, L.J. Guibas, P. Hanrahan, Q.-X. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, F. Yu, ShapeNet: an information-rich 3D model repository, abs/1512.03012, 2015.

[22] B. Kerbl, G. Kopanas, T. Leimkuehler, G. Drettakis, 3D Gaussian splatting for real-time radiance field rendering, ACM Trans. Graph. (TOG) 42 (2023) 1–14.

[23] Q. Team, Qwen3 Technical Report, arXiv:2505.09388 2025.

[24] Y. Tang, X. Han, X. Li, Q. Yu, Y. Hao, L. Hu, M. Chen, MiniGPT-3D: efficiently aligning 3D point clouds with large language models using 2D priors, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 6617–6626.

[25] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, J. Lu, Point-BERT: pre-training 3D point cloud transformers with masked point modeling, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 19291–19300.

[26] Z. Qi, R. Dong, G. Fan, Z. Ge, X. Zhang, K. Ma, L. Yi, Contrast with reconstruct: contrastive 3D representation learning guided by generative pretraining, in: International Conference on Machine Learning, PMLR, 2023, pp. 28223–28243.

[27] A. Marah, A. Jyoti, B. Sebastien, C.T.M. Caio, C. Weizhu, D.G. Allie, E. Ronen, G. Sivakanth, G. Suriya, J. Mojan, K. Piero, T.L. Yin, L. Yuanzhi, N. Anh, D.R. Gustavo, S. Olli, S. Adil, S. Shital, S. Michael, S.B. Harkirat, T.K. Adam, W. Xin, W. Rachel, W. Philipp, Z. Cyril, Z. Yi, Phi-2: The surprising power of small language models, 2023.

[28] C. Zhu, T. Wang, W. Zhang, J. Pang, X. Liu, LLaVA-3D: a simple yet effective pathway to empowering LMMs with 3D-awareness,abs/2409.18125 2024.

[29] H. Liu, C. Li, Q. Wu, Y.J. Lee, Visual instruction tuning, Adv. Neural Inf. Process. Syst. 36 (2023) 34892–34916.

[30] Y. Hong, H. Zhen, P. Chen, S. Zheng, Y. Du, Z. Chen, C. Gan, D-LLM: injecting the 3D world into large language models, Adv. Neural Inf. Process. Syst. 3 (2023) 20482–20494.

[31] J. Huang, S. Yong, X. Ma, X. Linghu, P. Li, Y. Wang, Q. Li, S.-C. Zhu, B. Jia, S. Huang, An embodied generalist agent in 3D world, in: Proceedings of the 41st International Conference on Machine Learning, 2024.

[32] H. Huang, Z. Wang, R. Huang, L. Liu, X. Cheng, Y. Zhao, T. Jin, Z. Zhao, Chat-3d v2: bridging 3D scene and large language models with object identifiers, 2023. abs/2312.08168

[33] M. Abdin, et al., Phi-3 technical report: a highly capable language model locally on your phone, 2024. abs/2404.14219

[34] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, J. Lin, 2025. Qwen2.5-vl technical report. abs/2502.13923

[35] A. Dubey, et al., The Llama 3 Herd of Models,2024. abs/2407.21783

[36] Z. Wang, J. Lorraine, Y. Wang, H. Su, J. Zhu, S. Fidler, X. Zeng, LLaMa-Mesh: unifying 3D mesh generation with language models, 2024. abs/2411.09595

[37] N. Reimers, I. Gurevych, Sentence-BERT: sentence embeddings using siamese bert-networks, in: Conference on Empirical Methods in Natural Language Processing, 2019.

[38] T. Gao, X. Yao, D. Chen, Simcse, Simple contrastive learning of sentence embeddings, in: 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Association for Computational Linguistics, 2021, pp. 6894–6910.

[39] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Annual Meeting of the Association for Computational Linguistics, 2002.

[40] S. Banerjee, A. Lavie, METEOR: an automatic metric for mt evaluation with improved correlation with human judgments, in: IEEEvaluation@ACL, 2005.

[41] C.-Y. Lin, ROUGE: a package for automatic evaluation of summaries, in: Annual Meeting of the Association for Computational Linguistics, 2004.

[42] T. Luo, C. Rockwell, H. Lee, J. Johnson, Scalable 3D captioning with pretrained models, Adv. Neural Inf. Process. Syst. 36 (2023) 75307–75337.