IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

# GaussianVLM: Scene-centric 3D Vision-Language Models using Language-aligned Gaussian Splats for Embodied Reasoning and Beyond

Anna-Maria Halacheva[1], Jan-Nico Zaech[1], Xi Wang[2], Danda Pani Paudel[1], Luc Van Gool[1]

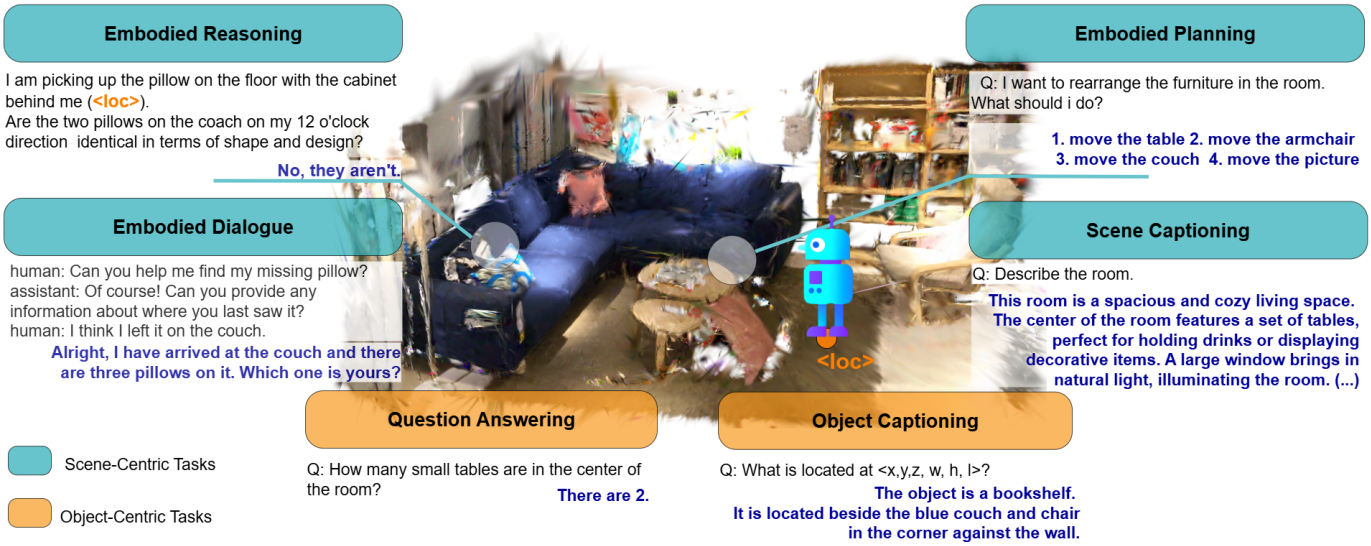


Fig. 1: The proposed GaussianVLM performs comprehensive scene understanding in natural language for 3D scenes represented as Gaussian Splats. It adopts a fully scene-centric approach, building a global, language-augmented scene representation. This enables effective handling of both scene- and object-level tasks – requiring multi-object reasoning, spatial understanding, global context, and fine-grained analysis – suitable for embodied reasoning and beyond.

***Abstract*—As multimodal language models advance, their application to 3D scene understanding is a fast-growing frontier, driving the development of 3D Vision-Language Models (VLMs). Current methods show strong dependence on object detectors, introducing processing bottlenecks and limitations in taxonomic flexibility. To address these limitations, we propose a scene-centric 3D VLM for 3D Gaussian splat scenes that employs language- and task-aware scene representations. Our approach directly embeds rich linguistic features into the 3D scene representation by associating language with each Gaussian primitive, achieving early modality alignment. To process the resulting dense representations, we introduce a dual sparsifier that distills them into compact, task-relevant tokens via task-guided and location-guided pathways, producing sparse, task-aware global and local scene tokens. Notably, we present the first Gaussian splatting-based VLM, leveraging photorealistic 3D representations derived from standard RGB images, demonstrating strong generalization: it improves performance of prior 3D VLM (LL3DA [8]) five folds, in out-of-the-domain settings.**

***Index Terms*—Semantic Scene Understanding; AI-Based Methods; Deep Learning for Visual Perception**

## I. Introduction

TO act intelligently in the physical world, embodied agents benefit from a rich, structured understanding of 3D scenes – capturing not only objects but also spatial context, relationships, and semantics [34], [40], [44], [45], [51]. Such scene understanding enables agents to move toward advanced tasks like embodied reasoning and planning, spanning multiple modalities [19], [29], [31], [32], [50] . While recent 3D VLMs have advanced towards addressing 3D vision-language tasks for embodied agents, they are predominantly object-centric, introducing a critical dependency on object detectors [8], [19], [21], [22], [60]. This creates a mismatch with the core objective of generic scene understanding, forcing models into predefined granularities, limited taxonomies, and neglecting global context and spatial relationships [16], [36]. In this work, we propose to shift from object-centric to scene-centric representations by embedding language features directly into the spatial structure of the environment. Each element of the 3D scene, represented either as a point or a Gaussian splat, is enriched with continuous language features, e.g. CLIP [38], SigLIP [56]. This allows us

Manuscript received: June, 7, 2025; Revised August, 25, 2025; Accepted September, 25, 2025.

This paper was recommended for publication by Editor Markus Vincze upon evaluation of the Associate Editor and Reviewers' comments.

This work was supported by the Ministry of Education and Science of Bulgaria (support for INSAIT, part of the Bulgarian National Roadmap for Research Infrastructure).

[1]First, Second, Fourth and Fifth Author are with INSAIT, Sofia University "St. Kliment Ohridski", Bulgaria name.surname@insait.ai

[2]Third Author is with INSAIT, Sofia University "St. Kliment Ohridski", Bulgaria, ETH Zurich, Switzerland, and TU Munich, Germany name.surname@inf.ethz.ch

Digital Object Identifier (DOI): see top of this page.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

to construct a language-aligned scene representation without relying on predefined object categories. Our scene-centric 3D VLM thus can answer complex questions related to both objects and scenes, as shown in Fig. 1.

However, directly embedding language features at the fine-granularity of the scene elements results in extremely dense representations in the tens of thousands tokens per scene. We argue that using the existing solutions, meaningfully understanding such representations via LLMs is a very challenging task – due to the high density of high-dimensional language features. To address this, we introduce a dual sparsifier module that efficiently utlizes dense language representations while preserving semantic fidelity. The dual nature of our sparsifier has two pathways: task-guided and location-guided. The task-guided sparsifier selects scene tokens based on global task relevance, and the location-guided sparsifier retrieves fine-grained features conditioned on spatial cues in the task, as shown in Fig. 2. The location-based sparsifier selects the language features of the Gaussians within the Region-of-Interest (ROI) around the location from the task, reducing them to a few ROI tokens. The task-guided sparsifier takes as input the dense scene tokens and the task tokens, using the latter in cross-attention to guide the sparsification process. As a result, the dense features are reduced to 128 task-selected scene tokens. The obtained sparse scene representation, consisting of the ROI tokens and task-selected tokens, is passed together with the task tokens to an LLM for response generation.

Finally, we develop the first 3D VLM operating on Gaussian Splatting (GS) that naturally fuses geometry and appearance information [26]. Note that unlike point clouds, Gaussian splats capture detailed 3D textures – in addition to the geometry – which is necessary for generic 3D scene understanding of our interest. For the more, with the recent developments the high-quality 3DGS can be realistically acquired using only RGB cameras. We demonstrate that our model, GaussianVLM, maintains strong task performance in real-world settings. We evaluate GaussianVLM and a state-of-the-art (SOTA) point-cloud based VLM [8] on an in-house question-answering (QA) task for counting objects in ScanNet++ scenes [54]. On the utilized out-of-domain ScanNet++ scene representations, derived from RGB images, the GS-based GaussianVLM outperforms the SOTA point cloud-based 3D VLM five folds in terms of accuracy (Tab. III).

We evaluate GaussianVLM on a comprehensive suite of 3D vision-language tasks spanning both scene-centric (Tab. I) and object-centric settings (Tab. II). Across the board, GaussianVLM achieves state-of-the-art performance, outperforming the SOTA baselines [8], [22] on every benchmark. Showing the advantages of scene-centrism, GaussianVLM significantly outperforms previous methods on embodied scene-centric tasks, e.g., embodied reasoning (SQA3D [31] 49.4% vs. 47.0% top-1 exact match) and substantially improving dialogue and planning metrics (e.g., +155.3 CIDEr in Embodied Planning [19]). Importantly, the detector-free GaussianVLM also excels on object-centric benchmarks, e.g., achieving improved object captioning on Nr3D [1] (+15.0 METEOR, +9.3 ROUGE).

Overall, this work makes the following contributions:

- We introduce a fully scene-centric 3D VLM that achieves SOTA results, without requiring any dependencies on object detectors, on benchmark datasets for reasoning tasks required for embodied vision and beyond.
- We propose a dual sparsification mechanism to efficiently distill dense language-augmented scenes into compact, task-relevant representations, suitable for LLMs.
- We present the first language-grounded 3D VLM directly operating on 3D Gaussian Splat representations.

## II. Related Work

### A. Scene-Level Reasoning of Embodied Agents

Early benchmarks in embodied question answering (EQA) [15], [50] pioneered tasks requiring agents to reason from egocentric observations, primarily focusing on situated, navigation-oriented challenges. Subsequent research expanded this scope to include multi-hop and commonsense reasoning [31], as well as embodied planning and dialogue tasks [19]. Early solutions adapted architectures like MCAN [55] and ClipBERT [23], with ScanQA [3] introducing 3D scene-grounded QA via explicit reconstructions. This progression has culminated in generalist 3D VLMs [8], [22], [60], [61] unifying 3D scene understanding, reasoning, and planning.

### B. 3D Scene Tokenization

For effective VLMs, 3D scene tokenization transforms complex geometry into language-processable, semantically rich representations. Two prevalent strategies exist:

**Object-Level Tokenization.** A common paradigm [21], [22], [25], [49], [58], [60], [10], [27] produces object-level tokens by detecting individual objects and independently encoding their point clouds. This method, while semantically intuitive, is limited by object detector performance and neglects vital scene context (e.g., room layout, walls).

**Region-Based Tokenization.** Another approach [59], [61] encodes the entire scene into per-point features, then groups these points into a fixed number of regions (e.g., via kNN [59] or graph-based segmentation [61]). Averaging features within these regions creates region-level tokens, capturing broader context at reduced granularity. However, this risks over-smoothing by collapsing diverse information into single tokens. Additionally, predefining the number of regions is challenging: too many can introduce irrelevant data and increase cost, while too few may lose fine-grained details.

In contrast, we introduce **language-guided scene tokenization**. This method dynamically re-tokenizes the scene based on linguistic input and per-point language features, producing tokens focused on regions most relevant to the current task.

### C. Vision-Language-Aligned 3D Scene Understanding

Integrating language into 3D scene understanding introduces challenges, particularly in (1) achieving effective cross-modal alignment [8], [22], and (2) ensuring semantically rich vision features [61].

**Text-Vision Alignment.** Prior work commonly aligns 3D visual features with language by projecting each modality independently into a shared embedding space [17], [21], [22],

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

[49]. However, this often results in weak alignment due to the largely separate processing of the two modalities. In line with 2D vision-language models [24], other approaches employ learnable query tokens that attend to both visual and textual features, separately, [8], [61], aiming for information fusion. Nevertheless, these query-based methods frequently refine visual features before language interaction, limiting the language's impact on the initial visual encoding. Critically, a shared limitation across these strategies is that the 3D encoder features are generated without incorporating any language or task-relevant semantic cues, ultimately leading to a shallow alignment [47]. Our approach, in contrast, ensures strong text-vision alignment by embedding language features directly into the fine-grained spatial structure of the 3D scene.

**Vision Feature Quality.** Recent efforts in 3D scene representation and sparsification have aimed to improve VLM performance by increasing the vision feature expressiveness. Many approaches leverage multi-modal visual data (2D images, point clouds, meshes) [19], [21], [61], [27], [33] for rich scene information, yet they are computationally intensive and architecturally complex, often also with inefficient, task-agnostic sparsification. Region highlighting techniques [8], [18], [59] attempt to emphasize key regions alongside a global scene representation, but the persistent use of dense global representations limits scalability and and attentional focus. We avoid these limitations by (a) using easy-to-obtain expressive language-aligned features [47] as our scene representation, and (b) generating all scene tokens conditioned on the task.

## III. Method

We introduce GaussianVLM, a 3D VLM for indoor scene understanding. Given a 3D scene represented as Gaussian splats and a natural language prompt, GaussianVLM fuses language and 3D vision at multiple stages to generate a textual response. Notably, GaussianVLM is the first to leverage Gaussian splats as the 3D scene representation, and function exclusively in the language space, achieving this object detector-free. GaussianVLM relies on three key innovations: (1) a language-aware Gaussian splatting backbone [26] that predicts language features for each Gaussian, enabling direct language-based alignment between the scene and the prompt; (2) a task-guided sparsifier module generating a sparse scene representation by performing task-aware re-tokenization of the dense 3D backbone output; and (3) a location-guided sparsifier module for detector-free extraction of Region-of-Interest (ROI) information. We detail the GaussianVLM and the sparsifier components in the subsequent sections.

### A. GaussianVLM

Unlike previous approaches that rely on purely visual representations, our method integrates a 3D transformer that produces inherently language-grounded vision features. Specifically, we adopt SceneSplat [26] as our 3D vision module. SceneSplat processes scenes represented via Gaussian splats and predicts a SigLIP2 [47] language feature for each Gaussian end-to-end. To sparsify the resulting dense language features with a task-awareness, we introduce a dual sparsifier module. The sparsifier takes as input the dense language features and outputs sparse task-aware tokens. The sparse scene tokens are projected from the SigLIP2 space into the LLM space via a single linear projection. The resulting vision tokens are then concatenated with the user task tokens, tokenized via the LLM's tokenizer, and input into a frozen LLM augmented with Low-Rank Adaptation (LoRA) [20]. The LLM autoregressively generates responses to the user query, conditioned jointly on both visual and textual context. GaussianVLM (OPT-1.3B [57] as LLM) has a size of 1.8B parameters out of which 19M are learnable.

**Training Objective.** Similarly to many VLM training protocols [22], [28], [58], we follow a two-stage training with alignment and fine-tuning phase. During the alignment phase we freeze the 3D backbone and LLM tokenizer, training the sparsifier modules and the transformer for textual alignment of the vision tokens. The LLM is adapted using LoRA. Both stages share a unified training objective. Following [6], [22], [39], we use a prefix language modeling, where the model is conditioned on an input prefix and trained to autoregressively generate the target continuation:

$$\mathcal{L}(\theta, \mathcal{B}) = -\sum_{\{s_{\text{prefix}}, s_{\text{gt}}\} \in B} \sum_{t=1}^{|s_{\text{gt}}|} \log p_\theta \left( s_{\text{gt}}^{(t)} \mid s_{\text{gt}}^{(<t)}, s_{\text{prefix}} \right), \quad (1)$$

with $\theta$ as the model parameters, $\mathcal{B}$ - a batch of samples of prefix input $s_{\text{prefix}}$ (task prompt and vision tokens), and ground truth response $s_{\text{gt}}$. $s_{\text{gt}}^{(t)}$ denotes the $t$-th token in the ground truth response sequence.

To enhance spatial grounding, we pre-train the task-guided sparsifier to understand 3D location features using an object captioning task. Given the 3D coordinates of a labeled object, the model is trained to generate a visual token embedding similar to the corresponding label token. The location is encoded through learnable Fourier embeddings (Eq. 3) and the label text – with the SigLIP-2 tokenizer. We pre-train with a one-sided contrastive objective [38], encouraging the output embedding of the task-guided sparsifier $s_i$ to match its corresponding label embedding $l_i$ and remain distant from all other labels $l_j$ ($j \neq i$):

$$\mathcal{L}\text{contrast} = -\log \frac{\exp(s_i^\top l_i / \tau)}{\sum j = 1^N \exp(s_i^\top l_j / \tau)}, \quad (2)$$

where $s_i$ is the output **s**cene token of the sparsifier for the $i$-th instance, $l_i$ is the SigLIP-2 embedding of the corresponding **l**abel, $N$ is the number of labels in the batch, and $\tau$ is a temperature hyperparameter which we set to 0.07.

### B. Dual Sparsifier

**Task-Guided Sparsification.** SceneSplat processes 3D Gaussians into a dense sequence of tokens (one per Gaussian). Following established practices [8], [9], [18], [59], sampling 40k Gaussians yields a corresponding 40k output tokens, originating from different SceneSplat decoder layers (specifically, 589, 2.4k, and 40k). To address the computational demands of the dense scene representation and prioritize task-relevant information, we introduce a novel task-guided sparsification

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

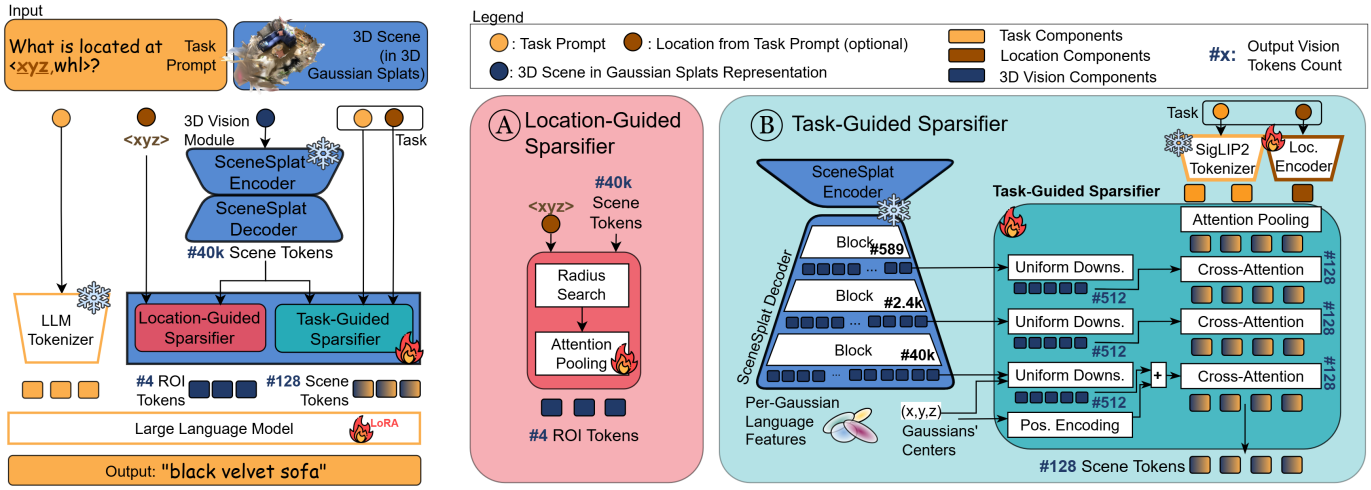


Fig. 2: The **GaussianVLM architecture** processes a user task prompt (query and optional location) and a 3D scene (Gaussian Splat representation). A 3D vision module (SceneSplat Transformer) predicts per-Gaussian language features. These dense features are then sparsified by a dual sparsifier module. The decoder's hidden states also inform the task-guided sparsifier. The dual sparsifier comprises: 1) a location-guided pathway that selects language features from Gaussians within a ROI around the task location, producing ROI tokens; and 2) a task-guided pathway that attends to dense scene tokens and SceneSplat decoder hidden states using task tokens (via cross-attention) to produce 128 task-selected scene tokens. The resulting sparse scene representation (ROI tokens + task-selected tokens), along with the task tokens, is input to an LLM for response generation.

module. This module leverages the language task to generate queries that guide the filtering of visual input via depth-wise cross-attention [5]. This process is applied iteratively to the output of each SceneSplat decoder layer, enabling a dynamic and context-aware reduction of visual information.

To mitigate the computational overhead of cross-attention on a large number tokens, we first uniformly downsample the representation to 512 tokens per decoder layer. Our ablation study (Sec. IV-G) confirms the sufficiency of this efficient approach, avoiding the need for more complex methods like kNN used in other models [18], [59]. Then, we further sparsify the representation to 128 tokens by performing cross-attention between the user's prompt (tokenized with SigLIP2) and these 3D features. All spatial locations $\text{loc}_{xyz} \in \mathbb{R}^3$ from the prompt are encoded via learnable Fourier embeddings [8]:

$$\text{pos}(\text{loc}_{xyz}) = [\sin(2\pi\, \text{loc}_{xyz} \cdot B)\ ;\ \cos(2\pi\, \text{loc}_{xyz} \cdot B)] \quad (3)$$

where $B \in \mathbb{R}^{3\times(d/2)}$ is a learnable matrix. If the prompt includes a bounding box, we extract $\text{loc}_{xyz}$ as its center. We apply attention pooling to the embeddings of the task tokens, generating a fixed set of 128 query vectors for cross-attention. Each SceneSplat decoder block has a corresponding cross-attention sparsifier block. The initial layer of these blocks performs cross-attention between the SceneSplat visual tokens and the task tokens. The subsequent layers process these intermediate visual features, further sparsifying and refining their semantic alignment with the language in a depth-wise manner. This process yields language-aware vision tokens that integrate global scene understanding with instance-level awareness. In the final sparsifier layer, we also inject positional information by encoding the center of the 512 downsampled Gaussian splats (Eq. 3) for location awareness.

**Location-Guided Sparsification.** For tasks that require a specific location, e.g., object captioning, we use an ROI magnifier. This module extracts features from a spherical region around the location provided in the prompt, using either a click's XYZ point or the center of a bounding box. We select neighboring points within a 15cm radius to focus on small objects. In empty ROI cases, we iteratively expand the radius by 15cm until points are captured. We then apply attention pooling to the language features of these selected points to generate 4 ROI tokens summarizing the region.

## IV. Experiments

### A. Dataset

We evaluate our model under the LL3DA, a SOTA 3D VLM, training protocol [8]. We also evaluate on embodied reasoning (SQA3D [31]), a popular 3D VLM benchmark [22], [61], where we follow the LEO [22] training protocol.

**LL3DA Training Protocol.** For the LL3DA protocol, we follow their one-stage joint training procedure. Training is performed on ScanRefer [7] (object captioning), ScanQA [3] (general question-answering (QA)), Nr3D [1] (object captioning), and the ScanNet subset of 3D-LLM [19] (diverse scene-centric tasks), focusing on multitask learning.

**LEO Training Protocol.** In the LEO setting, we adopt a two-phase training strategy with alignment and instruction tuning. To maintain compatibility with our scene-centric design and the LL3DA setup, we restrict training to the ScanNet subset of the LEO dataset. We align the visual and language modalities using the ReferIt3D dataset [1] providing detailed object captions. This phase helps the model ground linguistic features directly into the 3D scene representation. During the second stage, the model is further trained to follow natural

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

language instructions across multiple tasks using the SQA3D (situated QA) , ScanRefer, and ScanQA datasets.

### B. Tasks

We evaluate our model on a diverse set of 3D vision-language tasks drawn from the LL3DA and LEO benchmarks. We categorize the tasks into object-centric and scene-centric, reflecting differing demands on scene understanding.

**Object-Centric Tasks** require reasoning about discrete objects in the scene, often relying on explicit object annotations or localized queries. Those tasks include:

*Object Captioning.* We evaluate using ScanRefer [7] and Nr3D [1]. Given a natural language expression of an object and the object instance ID, we use the object's 3D bounding box center as the target location and prompt the model to generate a caption. To enhance linguistic diversity, we use GPT-4o to generate 40 syntactically and vocally varied paraphrases per prompt.

*Object-Centric QA.* We use ScanQA [3], which includes questions about object attributes, counts, and presence (e.g., "What is the color of the chair?"), targeting individual entities rather than spatial relationships or global context.

**Scene-Centric Tasks** require holistic reasoning about the environment, its layout, and the agent's situated context, without reducing the scene to individual object tokens. The *Situated Question Answering (SQA3D) [31]* requires the model to answer spatial or functional questions (e.g.,"What is on my left?") grounded in a given situational context (e.g., "I am washing my hands."), which demands an understanding of scene layout and affordances. For the *Embodied Planning [19]* task, the model generates high-level plans to complete tasks, leveraging the full scene structure to identify relevant objects and transitions. In *Scene Captioning [19]*, the model produces free-form descriptions summarizing the entire scene, requiring it to integrate geometry, object presence, and semantics into coherent language. *Embodied Dialogue [19]* introduces an interactive setting, where the model answers context-aware questions or participates in a dialogue about the scene, requiring dynamic grounding and multi-turn understanding.

### C. Metrics

For scene-centric tasks, where captions and answers typically encompass diverse and richly descriptive content, we report standard metrics including CIDEr [48], BLEU-4 [35], METEOR [4], ROUGE [13], exact-match accuracy, and Sentence-BERT [41] similarity. For object-centric tasks, we exclude BLEU-4 and CIDEr. BLEU, a precision-based metric, and CIDEr are overly sensitive to superficial n-gram overlap, rendering them unsuitable for evaluating long-form object captions [2]. Object captions, which extend beyond simple object naming to include rich context, can receive high scores even when the core referent is incorrect. This occurs because the two metrics reward overlapping phrases and frequent n-grams, even with an incorrect core referent. In contrast, we employ METEOR, ROUGE, and Sentence-BERT similarity, which offer superior handling of semantic alignment and partial matches [2], [41]. Specifically, METEOR incorporates synonym matching and alignment at the word and phrase level. ROUGE captures structural similarity and emphasizes recall without over-rewarding redundant context. Sentence-BERT directly evaluates semantic similarity in embedding space for robustness to paraphrasing.

### D. Implementation Details

Following prior work, each 3D scene is represented by 40k Gaussians sampled from the GaussianWorld [26] splats. We adopt OPT-1.3B [57] for LL3DA and Vicuna-7B [12] for LEO, following their respective training protocols. Both LLMs are float16-loaded for memory efficiency and fine-tuned using LoRa. The standard training protocol is applied: 5 alignment epochs and 10 instruction tuning epochs for LEO, and 32 epochs for LL3DA, completing in under one day on 8 A100-80 GPUs. Additionally, we pre-train our task-guided sparsifier for 5 epochs on object captioning. We use the AdamW optimizer with a weight decay of 0.1 and a cosine annealing learning rate schedule, decaying from ($10^{-4}$) to ($10^{-6}$). Evaluation is performed every 8 epochs for LL3DA and every epoch for LEO.

### E. Results and Analysis

The evaluation results (Tabs. I and II) confirm GaussianVLM's effectiveness across protocols and tasks. On the scene-centric SQA3D benchmark, GaussianVLM achieves an exact match accuracy of 49.4%, surpassing LEO's 47.0% by 2.4 percentage points. Under the LL3DA protocol, GaussianVLM significantly improves embodied dialogue (CIDEr: 145.9 $\rightarrow$ 270.1, +124.2) and planning (CIDEr: 65.1 $\rightarrow$ 220.4, +155.3), showcasing enhanced multi-object reasoning. In object-centric evaluations (Tab. II), our results are comparable (ScanQA) or superior (e.g., Nr3D METEOR 20.8 vs. 5.8) to existing methods, even without using object detectors.

### F. Real-World Generalization

To assess generalization to real-world data, we evaluate GaussianVLM and LL3DA using scene representations derived from RGB images. Unlike traditional point cloud-based VLMs, which often rely on laser-scanned geometry, our model is trained on photorealistic Gaussian splats, potentially offering better robustness to less structured inputs. For this out-of-domain (OOD) test, we use the ScanNet++ validation split [54] (not in our ScanNet training set). Specifically, we use GaussianWorld's [26] ScanNet++ scenes for our 3DGS representation, while the point cloud baseline (LL3DA) uses COLMAP [42], [43] reconstructions. To facilitate evaluation, we introduce a novel OOD object counting question-answering dataset (1000 Q/A pairs, automatically constructed from ScanNet++ segmentation annotations, excluding non-object categories). Across standard QA metrics (Exact Match, ROUGE, METEOR, CIDEr, Accuracy), our Gaussian splat-based model shows a significant performance advantage, outperforming the point cloud SOTA VLM (LL3DA) by 474% in accuracy on the GS scenes (Tab. III).

**IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.**

| | Embodied Dialogue | | | | | Embodied Planning | | | | | Scene Captioning | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Sim | C | B-4 | M | R | Sim | C | B-4 | M | R | Sim | C | B-4 | M | R |
| OPT-1.3B [57] | - | 0.31 | 0.23 | 5.62 | 4.83 | - | 0.16 | 0.13 | 0.24 | 3.56 | - | 0.0 | 0.84 | 8.40 | 11.7 |
| OPT-2.7B [57] | - | 0.38 | 0.39 | 7.38 | 6.28 | - | 0.10 | 0.26 | 3.59 | 4.35 | - | 0.11 | 0.00 | 6.60 | 12.32 |
| OPT-6.7B [57] | - | 0.25 | 0.43 | 6.88 | 6.16 | - | 0.00 | 0.28 | 3.65 | 3.94 | - | 0.06 | 1.13 | 8.99 | 16.96 |
| LLAMA-7B [46] | - | 0.27 | 0.50 | 7.81 | 6.68 | - | 0.04 | 0.29 | 3.53 | 4.71 | - | 0.2 | 0.92 | 7.00 | 12.31 |
| LL3DA* [8] | 48.2 | 145.9 | 22.2 | 40.9 | 36.7 | 50.2 | 65.1 | 7.1 | 20.8 | 32.2 | **66.4** | 0.2 | 3.0 | 19.4 | 18.4 |
| **GaussianVLM (Ours)** | **72.3** | **270.1** | **31.5** | **55.7** | **48.6** | **59.0** | **220.4** | **20.3** | **44.5** | **48.0** | 65.8 | **0.8** | **6.4** | **23.5** | **21.1** |

(a) LL3DA Scene-Centric Benchmarks. We compare 3D VLMs and frozen LLMs, following [8]. Our method, GaussianVLM, outperforms all baselines by a large margin.

| | SQA3D | | | | |
|---|---|---|---|---|---|
| | EM1 | C | B-4 | M | R |
| GPT3 [6] | 41.0 | - | - | - | - |
| ClipBERT [23] | 43.3 | - | - | - | - |
| SQA3D [31] | 46.6 | - | - | - | - |
| 3D-VisTA [60] | 48.5 | - | - | - | - |
| PQ3D [61] | 47.1 | - | - | - | - |
| LEO* [22] | 47.0 | 124.7 | 9.4 | 25.5 | 48.4 |
| **GaussianVLM (Ours)** | **49.4** | **129.6** | **17.1** | **26.4** | **50.2** |

(b) LEO Scene-Centric Benchmarks

TABLE I: Evaluation of SOTA 3D VLMs on **scene-centric** 3D vision-language tasks. (a) Results on the scene-centric benchmarks from LL3DA. (b) Results on the scene-centric benchmarks from LEO. We report results from specialist models (top) and generalist 3D VLMs (bottom). (*): reproduced. Evaluation metrics include CIDEr (C), BLEU-4 (B-4), METEOR (M), ROUGE (R), Sentence Similarity (Sim), and Top-1 Exact Match (EM1).

| | ScanRefer | | | ScanQA | | | Nr3D | | |
|---|---|---|---|---|---|---|---|---|---|
| | Sim | M | R | EM1 | M | R | Sim | M | R |
| Scan2Cap [11] | - | 21.4 | 43.5 | - | - | - | - | - | - |
| VoteNet+ MCAN [55] | - | - | - | 17.3 | 11.4 | 29.8 | - | - | - |
| ScanQA [3] | - | - | - | - | 13.14 | 33.3 | - | - | - |
| 3D-LLM [19] | - | 13.1 | 33.2 | **19.3** | 13.8 | 34.0 | - | - | - |
| 3D-VLP [53] | - | - | - | - | 13.5 | 34.5 | - | - | - |
| Scene-LLM [17] | - | 21.8 | 45.6 | - | 15.8 | - | - | - | - |
| LL3DA* [8] | 55.9 | 51.6 | 54.8 | 14.3 | 22.8 | 34.7 | 48.1 | 5.8 | 9.9 |
| **GaussianVLM (Ours)** | **59.1** | **52.4** | **57.4** | 14.4 | **22.9** | **34.8** | **48.2** | **20.8** | **19.2** |

TABLE II: Evaluation on **object-centric** LL3DA benchmarks. We report specialist models (top), and 3D VLMs (bottom). (*): reproduced. Models focusing on grounding (3D-LLM, 3D-VLP, Scene-LLM) and specialists were not reproduced due to differing objectives. Metrics: METEOR (M), ROUGE (R), Sentence Similarity (Sim), Top-1 Exact Match (EM1).

### *G. Ablation Study*

We conducted ablation studies in three categories (Tab. IV): [A] input ablations, [B] dual sparsifier architecture ablations, and [C] task sparsifier ablations. Input ablations confirm GaussianVLM's strong reliance on 3D input for accurate query answering. Furthermore, GaussianVLM on a Gaussian scene representation also outperforms its point cloud equivalent, especially on object-centric tasks, a result we attribute to Gaussian splats' ability to better preserve the fine-grained visual details. Our analysis further reveals that GaussianVLM's superior results are primarily due to: (a) the task-guided sparsifier, which leverages global context to provide task-conditioned scene-level awareness, and (b) the location-guided sparsifier, which offers localized information crucial for object-centric tasks. Removing either module results in a substantial performance drop (Tab. IV [B]). We further investigated the architecture of the task-guided sparsifier.

**Task-Guided Sparsifier.** We first examined the impact of task guidance. Replacing text-prompt-based queries with learnable queries caused a substantial performance decrease, especially for scene-centric tasks (Tab. IV), where the varied nature of prompts necessitates dynamic and task-aware selection of diverse visual cues. Next, ablating our depth-wise sparsification strategy by using only the final SceneSplat output (instead of intermediate features) caused a significant performance drop, particularly for scene-centric tasks requiring global context from earlier decoder layers. Finally, comparing our uniform downsampling to an advanced language-unaware alternative (attention pooling/k-NN) yielded no performance improvement, confirming the efficiency and information retention of our simpler approach.

## V. Limitations

While GaussianVLM demonstrates strong performance in 3D scene understanding, some limitations remain. First, relying on the 3D Gaussian Splatting representation introduces an additional reconstruction step, adding complexity compared to raw point-cloud methods. However, previous work and industrial applications have shown that this step is feasible on a scale [14], [26], [30], [37]. Second, performance can be sensitive to reconstruction quality; blur or reduced surface fidelity may affect small objects or surface-sensitive tasks. However, our evaluations indicate that GaussianVLM is robust, outperforming baselines on small-object captioning and appearance-based questions (Tab. V). Furthermore, GaussianVLM is not evaluated for interactive tasks, which we identify as an important direction for future work.

**IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.**

| Model | Accuracy (%) | Exact Match | CIDEr | METEOR | ROUGE |
|---|---|---|---|---|---|
| LL3DA [8] | 4.2 | 1.5 | 54.4 | 25.5 | 26.8 |
| **GaussianVLM (Ours)** | **24.1** | **9.3** | **120.0** | **35.2** | **47.3** |
| Improvement % | +474.0% | +520.0% | +120.6% | +38.0% | +76.5% |

TABLE III: Evaluation of QA on object counts on the out-of-domain ScanNet++ validation scenes.

| | | **Scene-Centric Tasks** | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **Embodied Dialogue** | | | | | **Embodied Planning** | | | | | **Scene Captioning** | | | | |
| | | Sim | C | B-4 | M | R | Sim | C | B-4 | M | R | Sim | C | B-4 | M | R |
| **[A] Input Ablation** | (1) Text-Only Input | 13.5 | 0 | 0 | 1.1 | 0 | 15.7 | 0 | 0 | 0.4 | 0 | 0.3 | 0 | 0 | 0.9 | 0.4 |
| | (2) 3D-Only Input | 34.7 | 67.5 | 8.8 | 24.9 | 19.9 | 37.0 | 46.6 | 4.1 | 21.1 | 25.8 | 37.8 | 0 | 0 | 0.2 | 0.3 |
| | (3) Point-Cloud Repr. | 71.9 | 267.4 | 30.7 | **55.9** | 48.3 | 58.4 | 218.4 | 20.0 | **44.5** | 47.1 | 65.7 | 2.0 | 4.8 | 23.0 | **21.1** |
| **[B] Dual Sparsifier** | (4) No Task-Guid. Spars. | 69.3 | 234.9 | 28.0 | 52.0 | 45.3 | 54.1 | 156.1 | 3.9 | 36.9 | 40.1 | 61.5 | 0.7 | 1.3 | 15.4 | 17.4 |
| | (5) No Loc.-Guid. Spars. | 68.9 | 233.4 | 28.1 | 52.0 | 44.9 | 56.8 | 195.0 | 12.0 | 41.0 | 44.8 | 63.4 | **2.5** | 3.1 | 19.6 | 20.9 |
| **[C] Task-Guid. Spars.** | (6) Learnable Queries | 71.4 | 267.0 | 31.3 | 55.5 | 48.5 | 58.2 | 218.5 | 17.7 | 44.2 | 47.8 | 59.6 | 0.1 | 1.6 | 15.1 | 17.9 |
| | (7) No Depth-Wise CA | 71.2 | 269.1 | 30.9 | 55.2 | 48.3 | 58.3 | 209.3 | 18.6 | 44.2 | 47.9 | 64.4 | 2.4 | 4.9 | 21.8 | **21.1** |
| | (8) kNN Sparsification | 71.2 | 261.6 | 31.1 | 54.9 | 47.8 | 58.0 | 218.0 | 17.1 | 44.2 | 47.8 | 63.3 | 1.7 | 5.4 | 22.0 | 20.0 |
| | **GaussianVLM (Ours)** | **72.3** | **270.1** | **31.5** | 55.7 | **48.6** | **59.0** | **220.4** | **20.3** | **44.5** | **48.0** | **65.8** | 0.8 | **6.4** | **23.5** | **21.1** |

| | | **Object-Centric Tasks** | | | | | |
|---|---|---|---|---|---|---|---|
| | | **ScanQA** | | | **Nr3D** | | |
| | | EM1 | M | R | Sim | M | R |
| **[A]** | (1) Text-Only Input | 0 | 1.6 | 0 | 32.0 | 10.2 | 9.6 |
| | (2) 3D-Only Input | 10.1 | 14.4 | 23.5 | 44.8 | 19.6 | 17.7 |
| | (3) Point-Cloud Repr. | 13.2 | 22.0 | 33.6 | 43.5 | 19.7 | 18.5 |
| **[B]** | (4) No Task-Guid. Spars. | **15.4** | 20.6 | 32.1 | 44.3 | 20.3 | 18.7 |
| | (5) No Loc.-Guid. Spars. | 14.2 | 21.5 | 34.2 | 44.1 | 19.0 | 18.9 |
| **[C]** | (6) Learnable Queries | 13.6 | 22.2 | 33.5 | 48.2 | **20.8** | 19.1 |
| | (7) No Depth-Wise CA | 13.9 | 22.4 | 33.9 | 47.9 | **20.8** | 19.0 |
| | (8) kNN Sparsification | 14.3 | **23.9** | **35.8** | **48.8** | **20.8** | 18.7 |
| | **GaussianVLM (Ours)** | 14.4 | 22.9 | 34.8 | 48.2 | **20.8** | **19.2** |

TABLE IV: Ablation Study of GaussianVLM. **[A] Input Ablation**: (1) Only text prompt provided, no 3D scene, (2) Only 3D scene provided, no task prompt. (3) Point cloud scene representation and point transformer (PTv3 [52]). **[B] Dual Sparsifier Architecture Ablation:** (4) No task-guided sparsifier, (5) No location-guided sparsifier, **[B] Task-Guided Sparsifier Architecture Ablation:** (6) Task prompt-based queries replaced with task-unaware learnable queries, (7) The three blocks of cross-attention (CA) are applied only to the final decoder output, not to hidden states, (8) Uniform downsampling replaced with a kNN and attention pooling strategy. **Evaluation metrics:** CIDEr (C), BLEU-4 (B-4), METEOR (M), ROUGE (R), Sentence Similarity (Sim), and Top-1 Exact Match (EM1).

| Method | ScanQA (Appear.) | | | ScanRefer (Small Obj.) | | |
|---|---|---|---|---|---|---|
| | EM | R | M | Sim | R | M |
| LL3DA | 0.29 | 0.43 | 0.23 | 0.38 | 0.29 | 0.25 |
| GaussianVLM | **0.32** | **0.47** | **0.24** | **0.51** | **0.42** | **0.36** |

TABLE V: Performance comparison on ScanQA (Appearance) and ScanRefer (Small Objects) tasks. **Evaluation metrics:** METEOR (M), ROUGE (R), Sentence Similarity (Sim), and Top-1 Exact Match (EM1).

## VI. Conclusion

We introduced GaussianVLM, a 3D VLM utilzing language-aligned Gaussian splats. With GaussianVLM, we proposed a paradigm shift in 3D vision-language understanding by moving away from object-centric representations towards a holistic, scene-centric and language-based approach. By directly embedding language features into the spatial structure of 3D scenes, GaussianVLM, overcomes the inherent limitations of object detector dependencies, enabling a more natural and comprehensive understanding of complex environments. We also proposed a dual sparsification module that effectively tackles the challenge of dense language-augmented scenes. The task-guided component distills the representation into compact, task-relevant features through task-guided selection on global context. Notably, with GaussianVLM, we presented a pioneering 3D VLM operating on Gaussian Splats, leveraging their rich geometric and appearance information for enhanced scene understanding/reasoning tailored to the embodied vision and beyond. Our extensive evaluations across a diverse suite of 3D vision-language tasks demonstrate the clear advantages of our scene-centric approach. GaussianVLM consistently achieves state-of-the-art performance, significantly outperforming existing methods on scene-centric tasks and also exhibiting strong results on object-centric benchmarks despite being detector-free. Finally, we empirically validated the practical generalization of our method, showing its improved performance on 3D data collected with more readily available equipment.

## References


[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. ReferIt3D: Neural listeners for fine-grained 3d object identification in real-world scenes. *ECCV*, 2020.

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. *ECCV*, 2016.

[3] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. *CVPR*, 2022.

[4] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. *ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005.

[5] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, et al. $\pi$0: A vision-language-action flow model for general robot control. *arXiv:2410.24164*, 2024.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, et al. Language models are few-shot learners. *NeurIPS*, 33, 2020.

[7] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. *ECCV*, 2020.

[8] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, et al. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. *CVPR*, 2024.

[9] Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, et al. End-to-end 3d dense captioning with vote2cap-detr. *CVPR*, 2023.

[10] Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, et al. Grounded 3d-llm with referent tokens. *arXiv:2405.10370*, 2024.

[11] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. *CVPR*, 2021.

[12] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 2023-04-14)*, 2(3), 2023.

[13] Lin Chin-Yew. Rouge: A package for automatic evaluation of summaries. *Proceedings of the Workshop on Text Summarization Branches Out, 2004*, 2004.

[14] Eric Cornwell, Dario Macangano, Stanford Lee, and Daniel Zilberman. Guidance for open source 3d reconstruction toolbox for gaussian splats on aws. *See https://aws-solutions-library-samples.github.io/compute/open-source-3d-reconstruction-toolbox-for-gaussian-splats-on-aws.html (accessed 2025-08-15)*, 2024.

[15] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, et al. Embodied question answering. *CVPR*, 2018.

[16] Alexandros Delitzas, Ayca Takmaz, Federico Tombari, Robert Sumner, Marc Pollefeys, et al. SceneFun3D: Fine-Grained Functionality and Affordance Understanding in 3D Scenes. *CVPR*, 2024.

[17] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual reasoning. *WACV*, 2025.

[18] Mei Guofeng, Lin Wei, Riz Luigi, Wu Yujiao, Poiesi Fabio, and Wang Yiming. Perla: Perceptive 3d language assistant. *CVPR*, 2025.

[19] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, et al. 3d-llm: Injecting the 3d world into large language models. *NeurIPS*, 2023.

[20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 2022.

[21] Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. *NeurIPS*, 2024.

[22] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, et al. An embodied generalist agent in 3d world. *ICML*, 2024.

[23] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, et al. Less is more: Clipbert for video-and-language learning via sparse sampling. *CVPR*, 2021.

[24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ICML*, 2023.

[25] Mingsheng Li, Xin Chen, Chi Zhang, Sijin Chen, Hongyuan Zhu, et al. M3dbench: Towards omni 3d assistant with interleaved multi-modal instructions. *ECCV*, 2025.

[26] Yue Li, Qi Ma, Runyi Yang, Huapeng Li, et al. Scenesplat: Gaussian splatting-based scene understanding with vision-language pretraining. *arXiv:2503.18052*, 2025.

[27] Xiongkun Linghu, Jiangyong Huang, Xuesong Niu, Xiaojian Ma, et al. Multi-modal situated reasoning in 3d scenes. *NeurIPS*, 2024.

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023.

[29] Qi Lv, Hao Li, Xiang Deng, Rui Shao, et al. Robomp2: A robotic multimodal perception-planning framework with mutlimodal large language models. *ICML*, 2024.

[30] Mengjiao Ma, Qi Ma, Yue Li, Jiahuan Cheng, et al. Scenesplat++: A large dataset and comprehensive benchmark for language gaussian splatting. *arXiv:2506.08710*, 2025.

[31] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, et al. Sqa3d: Situated question answering in 3d scenes. *ICLR*, 2023.

[32] Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. Situational awareness matters in 3d vision language reasoning. *CVPR*, 2024.

[33] Yunze Man, Shuhong Zheng, Zhipeng Bao, Martial Hebert, et al. Lexicon3d: Probing visual foundation models for complex 3d scene understanding. *NeurIPS*, 2024.

[34] Yang Miao, Iro Armeni, Marc Pollefeys, and Daniel Barath. Volumetric semantically consistent 3d panoptic mapping. *IROS*, 2024.

[35] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. *ACL*, 2002.

[36] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. *CVPR*, 2023.

[37] Pix4D. Transform your work with precise 3D scans and AR with PIX4Dcatch. *https://www.pix4d.com/product/pix4dcatch/ (accessed 2025-08-15)*, 2025.

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, et al. Learning transferable visual models from natural language supervision. *ICML*, 2021.

[39] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 2020.

[40] Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, et al. Sayplan: Grounding large language models using 3d scene graphs for scalable task planning. *CoRL*, 2023.

[41] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *EMNLP*, 2019.

[42] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. *CVPR*, 2016.

[43] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. *ECCV*, 2016.

[44] Ola Shorinwa, Johnathan Tucker, Aliyah Smith, Aiden Swann, et al. Splat-mover: Multi-stage, open-vocabulary robotic manipulation via editable gaussian splatting. *CoRL*, 2024.

[45] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *NeurIPS*, 2021.

[46] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*, 2023.

[47] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv:2502.14786*, 2025.

[48] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *CVPR*, 2015.

[49] Zehan Wang, Haifeng Huang, Yang Zhao, Ziang Zhang, and Zhou Zhao. Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes, 2023.

[50] Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. Embodied question answering in photorealistic environments with point cloud perception. *CVPR*, 2019.

[51] Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, et al. Tidybot: Personalized robot assistance with large language models. *IROS*, 2023.

[52] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, et al. Point transformer v3: Simpler, faster, stronger. *CVPR*, 2024.

[53] Dejie Yang, Zhu Xu, Wentao Mo, Qingchao Chen, Siyuan Huang, and Yang Liu. 3d vision and language pretraining with large-scale synthetic data. *arXiv:2407.06084*, 2024.

[54] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. *ICCV*, 2023.

[55] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. *CVPR*, 2019.

[56] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *ICCV*, 2023.

[57] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, et al. Opt: Open pre-trained transformer language models. *arXiv:2205.01068*, 2022.

[58] Yue Zhang, Zhiyang Xu, Ying Shen, Parisa Kordjamshidi, and Lifu Huang. Spartun3d: Situated spatial understanding of 3d world in large language models. *arXiv:2410.03878*, 2024.

[59] Hongyan Zhi, Peihao Chen, Junyan Li, Shuailei Ma, Xinyu Sun, et al. Lscenellm: Enhancing large 3d scene understanding using adaptive visual preferences. *arXiv:2412.01292*, 2024.

[60] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, et al. 3d-vista: Pre-trained transformer for 3d vision and text alignment. *CVPR*, 2023.

[61] Ziyu Zhu, Zhuofan Zhang, Xiaojian Ma, Xuesong Niu, et al. Unifying 3d vision-language understanding via promptable queries. *ECCV*, 2025.