# LLaVA$^3$ : Representing 3D Scenes Like a Cubist Painter to Boost 3D Scene Understanding of VLMs

**Doriand Petit**[1,2], **Steve Bourgeois**[1], **Vincent Gay-Bellile**[1], **Florian Chabot**[1], **Loïc Barthe**[2]

[1] Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
[2]IRIT, Université de Toulouse, CNRS, France
first.last@cea.fr, first.last@irit.fr

**Abstract**

Developing a multi-modal language model capable of understanding 3D scenes remains challenging due to the limited availability of 3D training data, in contrast to the abundance of 2D datasets used for vision-language models (VLMs). As an alternative, we introduce LLaVA³ (pronounced LLaVA Cube), a novel method that improves the 3D scene understanding capabilities of VLMs using only multi-view 2D images, and without requiring any fine-tuning. Inspired by Cubist painters, who represented multiple viewpoints of a 3D object within a single 2D picture, we propose to describe the 3D scene for the VLM through omnidirectional visual representations of each object. These representations are derived from an intermediate multi-view 3D reconstruction of the scene. Extensive experiments on 3D visual question answering and 3D language grounding show that our approach significantly outperforms previous 2D-based VLM solutions.

## 1 Introduction

3D scene understanding is a central goal in computer vision, with broad applications in areas such as robotics (Ni et al. 2023; Naseer, Khan, and Porikli 2018) and autonomous navigation (Guo et al. 2021). It involves a wide range of downstream tasks, from spatial reasoning and object decomposition to dense captioning and segmentation, covering both textual and pixel-level outputs, as illustrated in Figure 1.

Recent Vision-Language Models (VLMs) such as LLaVA (Liu et al. 2023b; Zhang et al. 2024) have demonstrated impressive capabilities in interpreting 2D images through multi-modal understanding and autoregressive language generation, enabling tasks like Visual Question Answering (VQA). Motivated by their success, multiple works have explored extending these models to 3D scene understanding. Specifically, 3D Multi-modal Large Language Models (3D MLLMs) are trained to directly process raw point clouds, whereas VLMs were extended to interpret multiple 2D images of the same scene at once. Yet, despite encouraging results, 3D performance still falls short of the rich reasoning capabilities achieved in 2D. Regarding VLMs, this gap is partly due to the inherent complexity of 3D reasoning, which requires integrating information across multiple viewpoints, spatial scales, and occlusions. Regarding 3D MLLMs, the main difficulty is related to the scarcity of 3D data to train such models.



In this paper, we propose to improve the 3D scene understanding ability of VLM from multi-view images by computing a novel visual 2D scene description that better captures the 3D nature of the scene. Inspired by works from ChatSplat (Chen, Wei, and Lee 2024) and SplatTalk (Thai et al. 2025), our approach first achieves a multi-view reconstruction of the scene in 3D, including a 3D field of LLaVA visual tokens. This 3D field is then used to provide 2D visual descriptions of the scene. However, we observe that those initial approaches sample tokens across the scene in a spatially unstructured manner, leading to redundant or inconsistent representations and limiting VLM performance.

To address this, we introduce a structured, object-centric approach inspired by the principles of Cubist art, which deconstructs 3D objects into 2D projections. We first segment the scene hierarchically into objects and compute an omnidirectional visual description for each object by sampling from the LLaVA field. These per-object token sets are semantically rich, spatially diverse, and context-aware, and can be fed as images to a frozen LLaVA model. Objects are ordered according to their spatial layout, providing a consistent and interpretable input for downstream reasoning.

Our evaluations demonstrate that LLaVA$^3$ outperforms other VLM-based solutions on 3D Visual Question Understanding (3D VQA) while allowing many 3D downstream tasks (Figure 1) without requiring any VLM fine-tuning.

To summarize, our contributions are:

1. We introduce the new concept of representing a 3D scene for a VLM as a collection of omni-directional visual-description of each of the 3D objects it contains.
2. We design the first objects hierarchy decomposition of a 3D Neural Field scene, facilitating the interpretation of the scene due to its discrete nature.
3. We propose to model jointly view-independent and view-dependent information contained in the high-dimension LLaVA visual tokens to better capture both objects semantics and their spatial relationships.
4. We demonstrate that LLaVA$^3$ outperforms other VLM-based solutions in term of 3D VQA and 3D grounding.

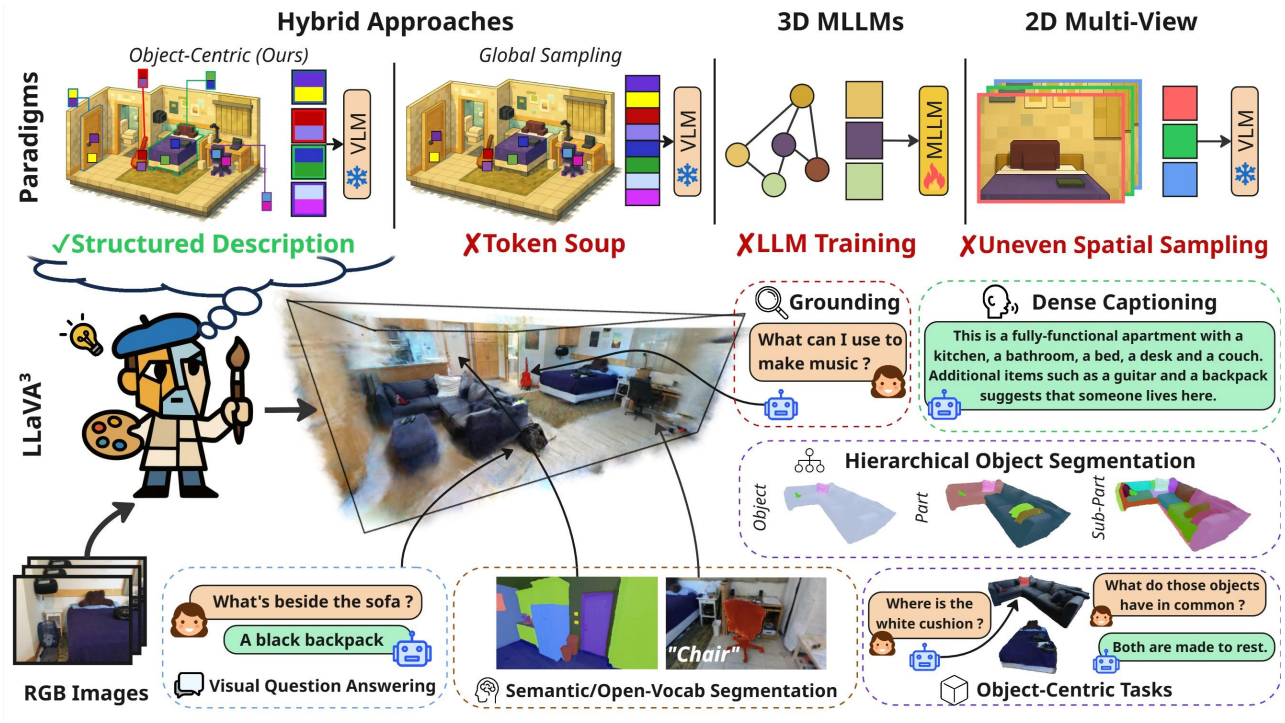


Figure 1: **LLaVA**$^3$ empowers 3D understanding ability of Vision Language Models (VLM) through a new paradigm of 2D visual representations of 3D scenes. Our representation relies on an object-centric description of the scene, each object being visually described from a multitude of viewpoints jointly. Reconstructed from multi-view images, this representation permits VLMs to achieve various tasks such as 3D Visual Question Answering, 3D Grounding or 3D Semantic Segmentation.

## 2 Related Works

**Large Multi-modal Models for 3D Scene Understanding.** Large Language Models (LLMs) such as GPT-4 (Achiam et al. 2023), Claude (Anthropic 2024), and LLaMA (Touvron et al. 2023) have established powerful foundations for natural language understanding, leading to the emergence of Large Multi-modal Models (LMMs) that extend language capabilities to visual domains. LLaVA (Liu et al. 2023b) (and its subsequent variants (Liu et al. 2023a, 2024; Li et al. 2024)) represents a seminal work in this direction, demonstrating that LLMs can understand and reason about visual content through a simple vision encoder-language model architecture connected by a projection layer.

Approaches to extends VLM abilities to 3D scene understanding can be grouped into several categories. First, 2D VLMs already exhibit impressive multi-view reasoning capabilities, often achieving surprisingly accurate 3D scene understanding from 2D inputs. However, these methods face key limitations: the restricted context window of language models and the mismatch between 2D image inputs and the inherently 3D nature of scenes. While recent works have sought to extend these architectures to 3D point clouds (Hong et al. 2023; Guo et al. 2023; Xu et al. 2024), adapting VLMs to point clouds remains difficult as point clouds are less informative than multi-view images and are harder to collect at scale, particularly for large-scale model training. A growing class of hybrid methods (Chen, Wei, and Lee 2024; Thai et al. 2025) addresses these issues by reconstructing the 3D scene and LLaVA feature field from multi-view images using NeRF or Gaussian Splatting, then summarizing the scene for the VLM via global sampling of the LLaVA feature field. This approach combines strengths of both paradigms: it leverages 3D reconstruction for dense spatial coverage while avoiding the need to train or fine-tune the VLM. Our work falls within this category, but we argue that existing sampling strategies, which usually consist in unstructured sampling across the scene, are sub-optimal. Instead, we decompose the scene into objects and construct per-object feature sets tailored for VLM consumption.

**NeRF and Feature Fields.** Neural Fields (Mildenhall et al. 2021) represent scenes using 3D feature grids (Müller et al. 2022) and neural networks. They are learnt from multi-view images and associated camera poses. Recent works have distilled pretrained image encoders into 3D feature fields to enable open-vocabulary understanding and semantic segmentation. Methods like LeRF (Kerr et al. 2023), OpenNeRF (Engelmann et al. 2024) or Decomposing-NeRF (Kobayashi, Matsumoto, and Sitzmann 2022) ground open-vocabulary embeddings (respectively CLIP (Radford et al. 2021), OpenSeg (Ghiasi et al. 2022) and LSeg (Li et al. 2022)) inside NeRF models, allowing textual querying across the scene. Other approaches leverage the Segment Anything Model (SAM) (Kirillov et al. 2023): SAM-NeRF projects SAM features into the scene for easy novel view

segmentation, while Garfield (Kim et al. 2024) rather proposes a continuous multi-scale scene decomposition by distilling masks using contrastive learning. Inspired by these ideas, LLaVA$^3$ learns two complementary 3D feature fields. The first one is aligned with LLaVA for semantic reasoning. Unlike previous works, it reconstructs the view-dependency of the LLaVA token to better capture the spatial relationships among the scene elements. The second field is aligned with SAM masks and CLIP features for scene decomposition. Unlike Garfield, our decomposition provides a discrete objects hierarchy that is easier to exploit than continuous volume-based decomposition.

# 3 Method

In this paper, we propose to improve the VLM ability to interpret 3D scenes from multiple views without any VLM fine-tuning. Instead of directly providing the visual-description of these multiple 2D views to the VLM, our solution provides an omni-directional visual-description for each object of the scene. As illustrated in Figure 2, our process first achieves a 3D grid-based NeRF reconstruction of the scene, including a LLaVA feature field (section 3.2). The reconstruction is then hierarchically decomposed into object/part/sub-parts, resulting into an explicit 3D object decomposition graph (section 3.3). An omnidirectional 2D visual-description of each object is computed by sampling tokens from the LLaVA-field equally amongst the object sub-components. This visual description is also designed to capture the object semantics while preserving the maximum of contextual information related to its relationships with the rest of the environment. Those objects' omnidirectional visual-descriptions are then provided as image tokens to the VLM, following an order depending on the object position in the 3D scene (section 3.4).

## 3.1 Preliminaries

**Neural Fields** (Mildenhall et al. 2021) (NeRFs) are learnable neural networks (possibly coupled with multi-resolution feature hashgrids (Müller et al. 2022)) over-fitted to individual scenes, which output density $\sigma$ and color $c$ from any 3D position and view direction queries. A 2D pixel color $\hat{C}$ is recovered by sampling points along a ray cast from the corresponding posed image and compositing them via volume rendering: $\hat{C}(r) = \sum_{i=0}^{N-1} w_i c_i$, with $w_i = T_i(1 - exp(-\sigma_i \delta_i))$ (which we denote as the density weights) and $T_i = exp(-\sum_{j=0}^{i-1} \sigma_j \delta_j)$, $c_i$ is the color of sample $i$ and $\delta$ is the distance between consecutive samples. The scene is optimized by minimizing the MSE loss $\mathcal{L}_{rgb} = ||\hat{C}(r) - C(r)||^2$ between rendered and ground truth colors. Feature fields are trained similarly by replacing RGB color with d-dimensional features, optimizing the model via comparison between NeRF rendered features and feature maps from pre-trained image encoder. Multiple feature fields can be learnt on the same model, using one decoder per feature plugged into joint or separate sets of grids.
**LLaVA-OV** (Li et al. 2024) integrates a SigLip (Zhai et al. 2023) image encoder $g_\psi$ with an LLM $f_\phi$, connected via an MLP projector $p_\theta$ that maps visual features into the language embedding space. The projected visual tokens are concatenated with textual tokens and jointly processed by the language model to produce textual outputs.

## 3.2 Reconstructing a LLaVA Feature Field

Our first objective is to reconstruct a dense 3D representation of the scene, including both geometry and LLaVA visual embeddings. To this end, we extend usual NeRF approach (Müller et al. 2022) with an additional LLaVA field. Following standard practice, we extract LLaVA token maps from training images and distill them into a 3D feature field. Despite LLaVA's low-resolution feature maps ($27 \times 27$), NeRF's multi-view consistency enables recovery of higher-resolution features, as shown in LeRF (Kerr et al. 2023). Based on ChatSplat (Chen, Wei, and Lee 2024) and SplatTalk (Thai et al. 2025) insights, we supervise the field with post-projector features ($p_\theta \circ g_\psi$) rather than SigLIP features. This choice stems from the projector's high sensitivity to reconstruction noise in SigLIP features, resulting in unintelligible tokens, while supervising with post-projector outputs offers more stability by operating directly in the VLM's token space.

**Reconstruction of high-dimensional feature field.** LLaVA tokens pose a unique challenge: their high dimensionality (3584D), sparsity, and lack of normalization make them difficult to learn reliably within a compact feature field without losing their semantic integrity. SplatTalk, based on Gaussian Splatting, addresses this by training a scene-specific auto-encoder that compresses tokens into a lower-dimensional, normalized space to stabilize training. This step is necessary in GS, where high-dimensional tokens cannot be directly stored or rendered due to memory limitations (Qin et al. 2024; Shi et al. 2024). However, it introduces information loss through two stages of imperfect compression (first via the auto-encoder, then during field learning) and adds additional pre-training overhead. In contrast, because we operate within a NeRF representation, we are not bound by these limitations and can learn full-resolution LLaVA features end-to-end using an auto-decoder architecture, preserving their full expressivity. Specifically, the feature field is decoded into normalized lower-dim feature $f$ before being mapped to the full token dimensionality $t$.

**Reconstructing semantics and spatial relationships.** Visuo-language features, such as CLIP, SigLIP or LLaVA features, encode both the semantics of individual objects and their spatial relationships. Because each viewpoint offers only a partial observation of the scene, the spatial information captured in these features is inherently view-dependent, while the object-level semantics remain view-independent.
To our knowledge, existing multi-view 3D reconstruction methods for feature fields typically rely on view-independent modeling (i.e. assigning a single static embedding to each 3D point, regardless of the viewpoint). This approach effectively captures the object's semantics while averaging out viewpoint-specific spatial relationships. Such filtering is well-suited for tasks like semantic segmentation, where object identity alone suffices. However, this becomes questionable in the context of vision-language models (VLMs) used for 3D scene analysis. In such cases, av-

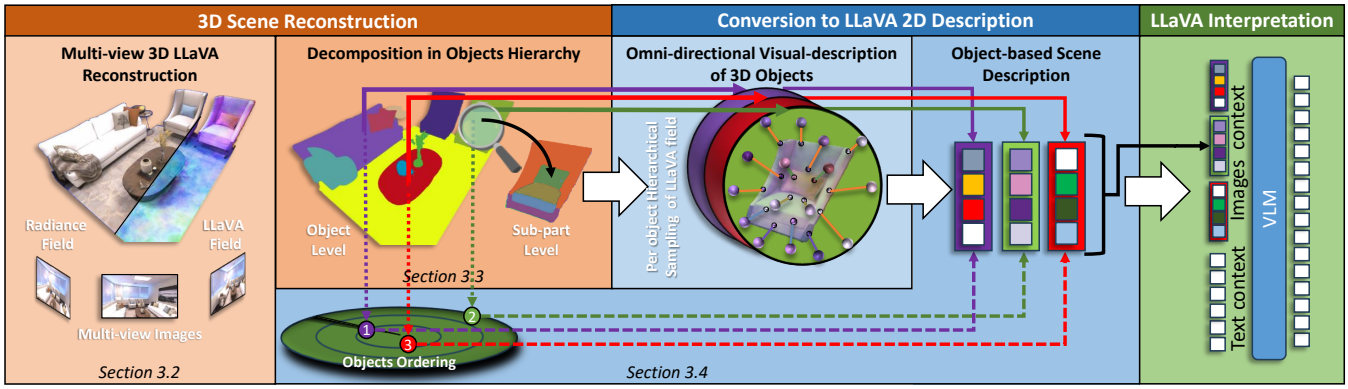


Figure 2: **Overview of LLaVA$^3$.** We first reconstruct the 3D scene as a NeRF from multi-view images with an associated LLaVA feature field. We also derive a hierarchical 3D segmentation of our NeRF. For each object, we create an omni-directional visual-description as a set of tokens. After object re-ordering, we can finally feed them to the VLM for 3D interpretation.

eraging view-dependent features can erase critical view-specific cues like inter-object relationships that are visible from very few input images. On the other hand, view-independent features benefit from multi-view aggregation, making them better at preserving object semantics.

To capture both semantics and spatial relationships, we jointly model in $f$ two complementary feature fields: one view-invariant $f_{VI}$ and one view-dependent $f_{VD}$, the latter being modeled as a deviation $\delta_{VD}$ of $f_{VI}$ to ensure consistency between these two:

$$f_{VD}(X, d) = f_{VI}(X) + \delta_{VD}(X, d)$$

with $X$ the 3D position in the scene and $d$ the direction of observation. For each training sample, the feature $f$ is first decoded into $f_{VD}$ and $f_{VI}$ (each one having its own decoder), which are decoded into full LLaVA tokens $t_{VD}$ and $t_{VI}$ using the same shared decoder and supervised independently with an MSE against the same ground-truth tokens.

### 3.3 Decomposing a NeRF into Objects Hierarchy

**Hierarchical NeRF Feature Field.** We draw inspiration from Garfield (Kim et al. 2024), which leverages SAM (Kirillov et al. 2023) masks across views to model implicitly a consistent 3D instance segmentation via a contrastive segment embedding field. In this setting, embeddings associated with rays falling within the same 2D mask are pulled together, while those from different masks are pushed apart, encouraging cross-view consistency. However, unlike Garfield's continuous spatial scale-space, which encodes scale in terms of spatial volume, we aim to construct a semantic scale-space that reflects discrete hierarchical levels: objects, parts, and sub-parts. This discrete structure aligns better with the goal of structured scene understanding and enables a clear representation of hierarchy.

We perform this by replacing Garfield's single scale-conditioned decoder with three separate decoders, all connected to a single set of feature grids and each dedicated to one level of the hierarchy. The training process is adapted accordingly: using the SAM discretization trick introduced in LangSplat (Qin et al. 2024), we separate the 2D SAM masks into three levels (object, part, subpart). Each set supervises the training of the corresponding decoder and feature field, enforcing a level-specific representation of the scene.

Finally, to both enhance the scene decomposition and enable additional scene understanding downstream tasks (Section 4.4), we further augment this field with a CLIP output. For each scale, an additional decoder predicts CLIP features from intermediate segment embeddings, trained using an MSE loss against CLIP embeddings computed from rendered SAM masks.

**3D Objects Hierarchy Extraction.** Now that we have learned a three-level hierarchical feature field, the next step is to construct a true hierarchical scene graph, that is, a tree structure in which any point sampled from the NeRF can be associated with a specific node in the objects hierarchy.

First, to derive full-scene segmentation from our hierarchical feature field, we cluster segment embbedings of a batch of randomly sampled rays using the clustering algorithm HDBScan (McInnes et al. 2017) independently at each semantic scale. To further encourage a scale-specific segmentation, we vary HDBScan's parameters per-scale (see supplementary material). Cluster centroids are computed as confidence-weighted embedding averages, such that any 3D point is assigned to its nearest centroid at each level.

However, this approach is not guaranteed to produce clean hierarchical segmentation, whether in terms of hierarchical misclassifications or reconstruction-based artifacts. To improve consistency, we introduce a three-fold refinement step that leverages CLIP features and the multi-scale hierarchy. For each HDBScan's segment, we compute a CLIP centroid and apply the following heuristics across scales: (i) discard noisy segments with high intra-cluster variance or low cardinal; (ii) split under-segmented regions when coarse clusters contain sub-segments with very dissimilar CLIP features, using the finer scale segmentation to compute new centroids; and (iii) merge over-segmented ones if feature centroids (either CLIP or SAM) are nearly identical.

To construct a hierarchical structure from our NeRF-based segmentation, we analyze the relationships between segments in a bottom-up manner. For each finer-level seg-

ment, we determine its parent by identifying the coarser-level segment it most frequently co-occurs with, using the HDBScan's rays clustering statistics. This process ensures a consistent and well-structured hierarchy in which each segment is uniquely assigned to a parent, resulting in a coherent and comprehensive decomposition of the scene.

### 3.4 Object-centric Description of the Scene

**Omnidirectional Visual-Description of a 3D Object.** For each object $i$, we aim to extract $N_f^i$ features that jointly describe its semantics and spatial relationships. To do so, we randomly cast rays through the supervision images and assign each resulting LLaVA feature to its corresponding object segment. This process continues until each object accumulates $N_f^i$ features. However, to construct an effective object representation, several questions must be addressed regarding feature selection, distribution, and quantity.

First, how can we ensure spatially balanced feature coverage across the whole object ? To address this, features are allocated regarding the objects hierarchy: first equally across part regions, then across sub-parts. This guarantees a uniform distribution over the components of the objects. Since the features are permutation-invariant, their internal ordering does not affect downstream processing.

A second key question concerns the type of features that should be selected: should they be purely view-independent (VI), purely view-dependent (VD), or a mixture of both ? VI features are better suited for object general semantics, while VD features excel at capturing spatial relationships that are inherently viewpoint-sensitive. We thus propose to distribute samples equally between VI and VD features. However, in tasks that involve spatial relationships from an observer's perspective (e.g. VQA), where references to "left" or "right" are common, we adopt an alternative strategy. We define a canonical viewing direction for each object and extract VD features only from rays that fall within a specified angular threshold of this direction, VI features being extracted otherwise. For simplicity, we define the canonical direction as the object most frequent viewing direction, computed via spherical binning over all ray directions used to observe it.

Finally, how many features should be allocated per object ? We fix the number of features per object to $N_f = W/O$, where $W$ is the maximal context window and $O$ is the number of objects. This strategy ensures uniform object representation. Ablations in the appendix confirm that this design leads to improved performance over variable-size allocations (depending on the object complexity for instance).

**Object-centric Scene Description.** The final step consists in aggregating those objects visual descriptions into a coherent scene-level representation, suitable for VLM reasoning. This process also raises important design questions, particularly around how to structure the sequence of object features in the prompt, given the constraints of using a frozen VLM.

Unlike 3D MLLM approaches that introduce positional encodings that require LLM fine-tuning, we explore how the arrangement of objects in the prompt affects how the VLM understands spatial relations across them.

To impose a consistent and layout-aware order, we adopt a radar-inspired sorting strategy. We first compute the 3D centroid of each object, then simulate a sweep originating from the scene center, sorting objects by their polar angle around the vertical axis. To resolve ties between objects with similar angles, we use a secondary term based on their radial distance from the center, with lower weight. This results in a deterministic and geometry-informed object ordering.

Finally, scene description is provided to the VLM as multi-view images, each object being tagged as a virtual image with a unique ID. This allows us to support grounding tasks: when prompted, the VLM outputs the most relevant ID for a given query, which can be mapped back to the corresponding segment in the NeRF reconstruction.

## 4 Experiments

### 4.1 Implementation Details

We implemented our method in the Nerfstudio (Tancik et al. 2023) framework. Every NeRF is trained with the Nerfacto model, a grid-based NeRF method coupled with several Mip-NeRF-360 (Barron et al. 2022) improvements. Each NeRF trained on ScanNet uses $80\%$ of the images (chosen based on estimated blurriness) for geometry and 400 selected images for the two feature fields. The selection was made by maximizing the scene coverage of the scene, as explained in the supplementary material. All our experiments (NeRF training and LLM inference) were run on one A100 GPU. Additional details on hyper-parameters, evaluation protocols and reproducibility can be found in the supp. mat., as well as additional ablative experiments.

### 4.2 3D Visual Question Answering (3D VQA)

**Datasets and Protocol.** We use ScanNet datasets (Dai et al. 2017) for evaluation, specifically the ScanQA (Azuma et al. 2022) validation set and the MSR3D (Linghu et al. 2024) test set. For ScanQA, we evaluate performance using standard n-gram-based metrics: CIDEr, METEOR, ROUGE, EM@1 and EM@1-Refined. For MSR3D, we adopt their GPT-based correctness score and follow their implementation with GPT-4o as the notation model.

**Qualitative Results.** Figures 3 and 1 showcase various example with different common errors solved by our method and object-centric VQA (i.e., feeding tokens of specific objects to the VLM rather than the whole scene).

**Results on ScanQA.** Similarly to SplatTalk evaluation, we compare our methods against several families of baselines on the ScanQA validation set and we report results of each in Table 1. "*3D LMMs*" lists several 3D large multi-modal models trained across a diverse range of tasks. However, some of those models use ScanQA train set, either during their training or their fine-tuning. Please note that all those models uses ScanNet and other indoor scenes as training dataset, thus limiting the generalization to other types of scenes and inducing a boosting bias in their performance evaluations. "*2D VLMs*" refers to models that only process multi-view images inputs. The object-based LLaVA-OV baseline is constructed by applying SAM to multi-view images and extracting LLaVA features from each resulting mask. These per-object features are then fed directly to the LLM. We build this baseline to evaluate the impact of using

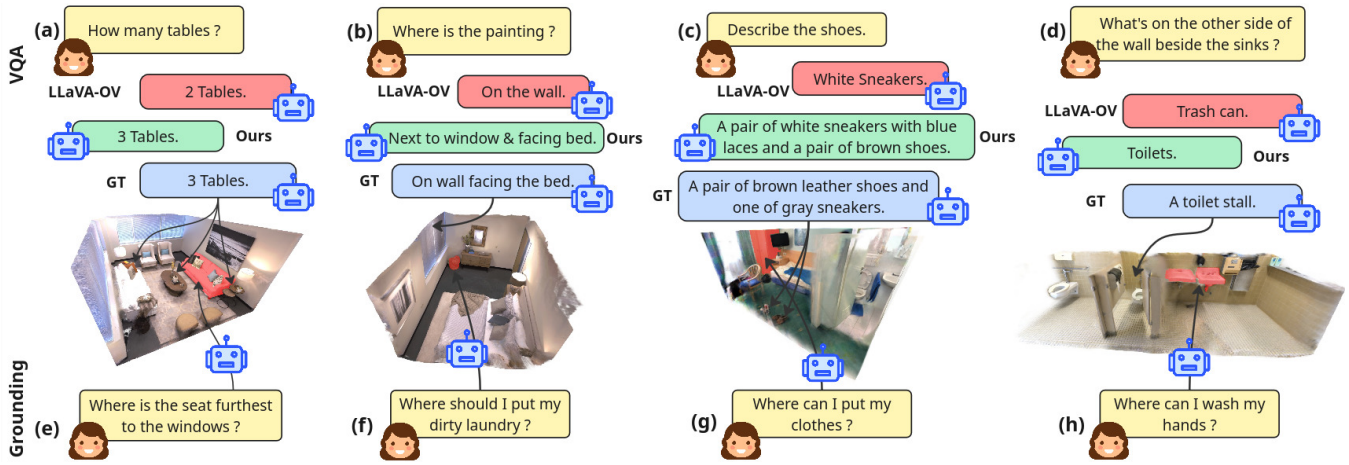


Figure 3: **Qualitative performance of our method on 3D VQA and Grounding**. By decomposing into objects the view-dependent features, our method avoids several very common issues in 3D VQA: (a) missing objects due to insufficient sampling, (b) weak inter-object spatial relationships, (c) loss of objects details and (d) bad multi-view understanding (i.e. relations between objects from different images). Our method can also perform grounding from different types of queries ((e),(f),(g),(h)).

| Method | Modality | CIDEr ↑ | METEOR ↑ | Rouge ↑ | EM@1 ↑ | EM@1-R ↑ |
|---|---|---|---|---|---|---|
| ***3D LMMs*** | | | | | | |
| 3D-VisTA* (T) | PC | 69.6 | 13.9 | 35.7 | 22.4 | - |
| Chat-3D-v2* (FT) | PC | 77.1 | 16.1 | 40.0 | - | - |
| LL3DA* (T) | PC | 76.8 | 15.9 | 37.3 | - | - |
| Scene-LLM* | PC + I | 80.0 | 15.8 | 35.9 | - | - |
| Scene-LLM* (FT) | PC + I | 80.0 | 16.6 | 40.0 | 27.2 | - |
| LEO* (T) | PC + I | 101.4 | 20.0 | 49.2 | 24.5 | 47.6 |
| ***2D VLMs*** | | | | | | |
| Claude* | I | 57.7 | 10.0 | 29.3 | - | - |
| GPT-4V* | I | 59.6 | 13.5 | 33.4 | - | - |
| LLaVA-NeXT-Video* | I | 46.2 | 9.10 | 27.8 | - | - |
| LLaVA-OV | I | 55.38 | 13.22 | 30.36 | 16.83 | 33.99 |
| LLaVA + Object-based | I | 51.78 | 12.87 | 29.46 | 16.05 | 32.45 |
| ***Hybrid Methods*** | | | | | | |
| SplatTalk* | I | 61.7 | 14.2 | 32.7 | 17.1 | 32.2 |
| SplatTalk (FT)* | I | 77.5 | 15.6 | 38.5 | 22.4 | 38.3 |
| Ray-NeRF | I | 60.56 | 13.82 | 31.56 | 20.83 | 32.77 |
| LLaVA$^3$ | I | **77.69** | **15.83** | **39.69** | **26.00** | **38.43** |

Table 1: **3D VQA Performance on ScanQA validation set.** FT and T refer to fine-tuning and training on ScanQA data. Grey entries denote incorporation of ground-truth objects as inputs. "*" refers to results taken from papers and bold indicates best performance among the image-based methods. PC refers to Point Cloud and I to Images.

| Methods | Counting | Existence | Attributes | Spatial | Navigation | Others | Overall |
|---|---|---|---|---|---|---|---|
| ***Baselines*** | | | | | | | |
| LEO* | 0.8 | 15.5 | 11.8 | 7.3 | 2.3 | 15.3 | 7.8 |
| LLaVA-OV | 19.20 | 35.00 | 28.48 | 23.36 | 12.50 | 39.94 | 27.87 |
| Object LLaVA-OV | 15.19 | 33.95 | 29.18 | 18.85 | 10.79 | 43.31 | 25.02 |
| SplatTalk* | 19.6 | 60.3 | 44.0 | **35.8** | **35.5** | 61.8 | 41.8 |
| Ray-NeRF | 21.25 | 52.35 | 36.79 | 30.54 | 15.40 | 53.29 | 30.54 |
| ***Ablatives*** | | | | | | | |
| + *Object-Based* | +4.19 | +14.32 | +9.31 | -1.86 | +9.22 | +9.69 | +6.85 |
| + *Multi-Scale* | +1.35 | +3.04 | +3.08 | +0.65 | +1.05 | +3.55 | +2.5 |
| + *Filtering* | +0.30 | +0.93 | +1.42 | +0.36 | +0.22 | +1.60 | +0.55 |
| + *Radar Sweeping* | +2.05 | +3.28 | -0.11 | +1.48 | +3.21 | +4.18 | +2.09 |
| 100% VI | 29.14 | 73.92 | 50.49 | 31.17 | 29.1 | 72.31 | 42.53 |
| 100% VD | 29.01 | 71.99 | 49.99 | 31.98 | 30.02 | 73.49 | 42.60 |
| 50% VI - 50% VD | 29.34 | 74.13 | 51.02 | 32.43 | 30.20 | 72.99 | 43.84 |
| ***LLaVA$^3$*** | **29.51** | **75.00** | **51.60** | 33.30 | 31.35 | **73.99** | **44.89** |

Table 2: **3D VQA performance per question type on MSR3D** test set using their correctness score (↑). *Ablatives* first display the relative difference to the previous line down from Ray-NeRF baseline (with 100% VI features) and then uses the full method while only modifying the VI/VD distribution (LLaVA$^3$ uses the Adaptive VI-VD distribution).

object-centric inputs from 2D images, in contrast to our proposed 3D object-based pipeline. Lastly, "*Hybrid*" methods only use 2D VLMs but first reconstruct in 3D the scene using NeRF or GS from multi-view images of the scene. Alongside SplatTalk (generalist and LoRA) and our method, we introduce a Ray-NeRF baseline that samples LLaVA visual tokens from random rays across the NeRF field, without object structuring. This baseline isolates the benefit of object-centric feature extraction by contrasting it with an unstructured use of the same NeRF representation.

Most importantly, our base method outperforms all zero-shot NeRF and GS models and 2D VLMs. This demonstrates that we successfully derive a more comprehensive and efficient scene representation for a VLM than traditional multi-view input and random ray sampling. Specifically, both NeRF methods (especially ours) achieve higher performance than LLaVA-OV although we use the same LLM and input feature maps, meaning that although using a NeRF to ensure the use of visual tokens across the whole scene helps (+5.18 CIDEr), adding the object-centric representation of LLaVA$^3$ benefits significantly more (+22.31 CIDEr). However, the 2D object-based baseline (−3.60 against LLaVA-OV) demonstrates that this object-centric decomposition is beneficial only for 3D representation. When comparing to other paradigms, we notice better or comparable results to most *3D LMMs* (except for LEO, our metrics range between

$-4.6\%$ and $+13.2\%$ of the other baselines with an average of $+2\%$ across all metrics), despite not using any 3D-specific information, nor training our VLM on ScanNet data.

**Results on MSR3D.** To evaluate the impact of our different contributions, we showcase quantitative results in Table 2 on the MSR3D dataset, which decomposes into multiple question types, allowing fine-grained understanding of what brings each contribution. In addition to the LLaVA-2D baselines, SplatTalk and LEO, we progressively build LLaVA$^3$ on the Ray-NeRF baseline by adding our contributions one by one to demonstrate their individual impact.

First, we analyze our contributions. Decomposing the scene into objects has the highest effect on overall performance ($+6.85$), as it ensures a balanced distribution of the sampling across the scene. More in-depth balancing of the features via hierarchical sampling further increases the results ($+2.5$); especially for questions related to visual appearance (e.g. existence and attributes with resp. $+3.04$ and $+3.08$), as it gives the model more exhaustive information on each object. As expected, improving the scene decomposition quality by adding filtering steps slightly improves the results globally. Ordering the scene description via our radar sweeping strategy also noticeably helps the VLM, with a $2.09$ score bump. Regarding the view dependency, although replacing VI features with VD features results in similar overall performance (but still different detailed distribution), combining both features in an even split improves the performance. It appears that VI features help the model to reason over descriptive queries (counting, existence, attributes), while the VD features help with spatial and navigation questions. Finally, the adaptive VI-VD used in LLaVA$^3$ is better than splitting evenly the features, as choosing one canonical view direction helps follow the observer's perspective.

Our full model outperforms most comparative baselines across the majority of categories. LEO, which is specifically trained for ScanQA-style queries, struggles to generalize to this dataset. The 2D baselines show overall lower performance, largely due to the small number of input images. Similarly, the other hybrid methods such as SplatTalk and our Ray-NeRF baseline under-perform on most categories, with respective drops of $-3.09$ and $-14.35$ points.

### 4.3 Grounding

**Datasets.** We evaluate grounding on ScanNet-based Sr3D+ and Nr3D, which provide natural language queries with ground-truth instance IDs. We use a subset of ScanNet scenes: 0011, 0030, 0046,0086, 0222, 0378, 0389 and 0435.

**Protocol.** Following standard baselines, we perform grounding by retrieving a binary segmentation point cloud from our segmented NeRF using the instance ID predicted by the LLM. To convert our NeRF into a point cloud, we adopt the OpenNeRF protocol: for each view, we render the segmentation mask and back-project it onto the point cloud. For each point, we average the instance class feature across views and use the centroids to determine the segmentation. We then compute standard detection metrics: a prediction is correct if the segmentation IoU exceeds a threshold. We report accuracy at 0.1 and 0.25 IoUs, following prior works.

**Results.** Results are reported in Table 3, with qualitative examples provided in Figure 3. First, our method heavily outperforms existing NeRF-based open-vocabulary baselines such as LeRF and OpenNeRF (resp. $+245\%$ and $+60\%$ in average). This improvement is expected, as those methods rely solely on CLIP-based similarity measures, which are effective when object names are explicitly mentioned in the query. In contrast, grounding queries are phrased as spatial indications, making such non-reasoning approaches less effective. The ConceptGraph baseline, which stores an explicit caption per-object, achieves lower performance compared to ours, highlighting the fact that our object-centric tokenization gives the LLM the ability to reason over objects without any retraining. Using our hierarchical representation, we can also perform sub-scale grounding, as illustrated in Figure 1.

| Methods | Sr3D++ | | Nr3D | |
|---|---|---|---|---|
| | Acc@0.1 ↑ | Acc@0.25 ↑ | Acc@0.1 ↑ | Acc@0.25 ↑ |
| LeRF | 6.88 | 1.97 | 5.53 | 1.43 |
| OpenNeRF | 9.07 | 4.03 | 9.70 | 5.11 |
| ConceptGraph | <u>13.3</u> | <u>6.2</u> | <u>16.0</u> | <u>7.2</u> |
| LLaVA$^3$ | **14.41** | **6.54** | **16.17** | **7.81** |

Table 3: **3D Grounding Performance** on Sr3D+ and Nr3D.

### 4.4 Semantic Segmentation and Other Segmentation Tasks

By leveraging the SAM-CLIP feature field (Section 3.3), we enable a broad range of segmentation tasks, a fundamental aspect of downstream 3D scene understanding. LLaVA$^3$ supports semantic segmentation (see Figure 1), but also open-vocabulary segmentation and instance segmentation, all within a hierarchical framework.

For semantic segmentation specifically, evaluation on standard benchmarks using both Replica and ScanNet scenes showcases that our method significantly outperforms other NeRF-based approaches such as LeRF (Kerr et al. 2023), OpenNeRF (Engelmann et al. 2024), and DiSCO-3D (Petit et al. 2025), achieving respectively up to +146% mIoU / +91% mAcc, +21% / +7%, and +15% / +14% gains. It can also be noted that it compares favorably with explicit scene-graph-based models, outperforming ConceptGraph (+36% mIoU and +20% mAcc) and matching HOV-SG (+4% and -3%). Additional results, figures, and analyses are provided in the supplementary.

## 5 Conclusion

We presented LLaVA$^3$, a novel object-centric approach that improves the 3D scene understanding ability of VLMs from multi-view images without fine-tuning. Inspired by Cubism, our method decomposes the scene into objects hierarchy and compute for each of them an omnidirectional 2D visual representation that captures both semantic and spatial relationships. Experimental results demonstrate that it enables VLMs to reason effectively over 3D content, avoiding common pitfalls such as object duplication or limited context windows, and outperforming other VLM-based solutions. LLaVA$^3$ enables a large variety of downstream tasks, including 3D VQA, grounding and semantic segmentation.

## Acknowledgements

This publication was made possible by the use of the CEA List FactoryIA supercomputer, financially supported by the Ile-de-France Regional Council.

## References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2024. Introducing the next generation of Claude.

Azuma, D.; Miyanishi, T.; Kurita, S.; and Kawanabe, M. 2022. ScanQA: 3D Question Answering for Spatial Scene Understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*

Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 5470–5479.

Chen, H.; Wei, F.; and Lee, G. H. 2024. ChatSplat: 3D Conversational Gaussian Splatting. *arXiv preprint arXiv:2412.00734*.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5828–5839.

Engelmann, F.; Manhardt, F.; Niemeyer, M.; Tateno, K.; Pollefeys, M.; and Tombari, F. 2024. OpenNeRF: Open Set 3D Neural Scene Segmentation with Pixel-Wise Features and Rendered Novel Views. *arXiv preprint arXiv:2404.03650*.

Ghiasi, G.; Gu, X.; Cui, Y.; and Lin, T.-Y. 2022. Scaling open-vocabulary image segmentation with image-level labels. In *Eur. Conf. Comput. Vis.*, 540–557. Springer.

Guo, Z.; Huang, Y.; Hu, X.; Wei, H.; and Zhao, B. 2021. A survey on deep learning based approaches for scene understanding in autonomous driving. *Electronics*, 10(4): 471.

Guo, Z.; Zhang, R.; Zhu, X.; Tang, Y.; Ma, X.; Han, J.; Chen, K.; Gao, P.; Li, X.; Li, H.; et al. 2023. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*.

Hong, Y.; Zhen, H.; Chen, P.; Zheng, S.; Du, Y.; Chen, Z.; and Gan, C. 2023. 3d-llm: Injecting the 3d world into large language models. *Adv. Neural Inform. Process. Syst.*, 36: 20482–20494.

Kerr, J.; Kim, C. M.; Goldberg, K.; Kanazawa, A.; and Tancik, M. 2023. LERF: Language Embedded Radiance Fields. In *Int. Conf. Comput. Vis.*

Kim, C. M.; Wu, M.; Kerr, J.; Goldberg, K.; Tancik, M.; and Kanazawa, A. 2024. Garfield: Group anything with radiance fields. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 21530–21539.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Int. Conf. Comput. Vis.*, 4015–4026.

Kobayashi, S.; Matsumoto, E.; and Sitzmann, V. 2022. Decomposing nerf for editing via feature field distillation. *Adv. Neural Inform. Process. Syst.*, 35: 23311–23330.

Li, B.; Weinberger, K. Q.; Belongie, S.; Koltun, V.; and Ranftl, R. 2022. Language-driven Semantic Segmentation. In *Int. Conf. Learn. Represent.*

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Li, Y.; Liu, Z.; and Li, C. 2024. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv preprint arXiv:2408.03326*.

Linghu, X.; Huang, J.; Niu, X.; Ma, X.; Jia, B.; and Huang, S. 2024. Multi-modal Situated Reasoning in 3D Scenes. In *Advances in Neural Information Processing Systems*, volume 37, 140903–140936. Curran Associates, Inc.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023a. Improved Baselines with Visual Instruction Tuning.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual Instruction Tuning. In *Adv. Neural Inform. Process. Syst.*

McInnes, L.; Healy, J.; Astels, S.; et al. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11): 205.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.*, 41(4): 102:1–102:15.

Naseer, M.; Khan, S.; and Porikli, F. 2018. Indoor scene understanding in 2.5/3d for autonomous agents: A survey. *IEEE access*, 7: 1859–1887.

Ni, J.; Chen, Y.; Tang, G.; Shi, J.; Cao, W.; and Shi, P. 2023. Deep learning-based scene understanding for autonomous robots: A survey. *Intelligence & Robotics*, 3(3): 374–401.

Petit, D.; Bourgeois, S.; Gay-Bellile, V.; Chabot, F.; and Barthe, L. 2025. DiSCO-3D : Discovering and segmenting Sub-Concepts from Open-vocabulary queries in NeRF. In *Int. Conf. Comput. Vis.*

Qin, M.; Li, W.; Zhou, J.; Wang, H.; and Pfister, H. 2024. Langsplat: 3d language gaussian splatting. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 20051–20060.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Shi, J.-C.; Wang, M.; Duan, H.-B.; and Guan, S.-H. 2024. Language embedded 3d gaussians for open-vocabulary scene understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 5333–5343.

Tancik, M.; Weber, E.; Ng, E.; Li, R.; Yi, B.; Wang, T.; Kristoffersen, A.; Austin, J.; Salahi, K.; Ahuja, A.; et al.

2023. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, 1–12.

Thai, A.; Peng, S.; Genova, K.; Guibas, L.; and Funkhouser, T. 2025. SplatTalk: 3D VQA with Gaussian Splatting. *arXiv preprint arXiv:2503.06271*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Xu, R.; Wang, X.; Wang, T.; Chen, Y.; Pang, J.; and Lin, D. 2024. PointLLM: Empowering Large Language Models to Understand Point Clouds. In *Eur. Conf. Comput. Vis.*

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Int. Conf. Comput. Vis.*, 11975–11986.

Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2024. Vision-language models for vision tasks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*