SplatTalk: 3D VQA with Gaussian Splatting

Anh Thai[1,2*] Songyou Peng[2] Kyle Genova[2] Leonidas Guibas[2] Thomas Funkhouser[2]

[1]Georgia Institute of Technology [2]Google DeepMind

[*]Work done as a student researcher at Google DeepMind.

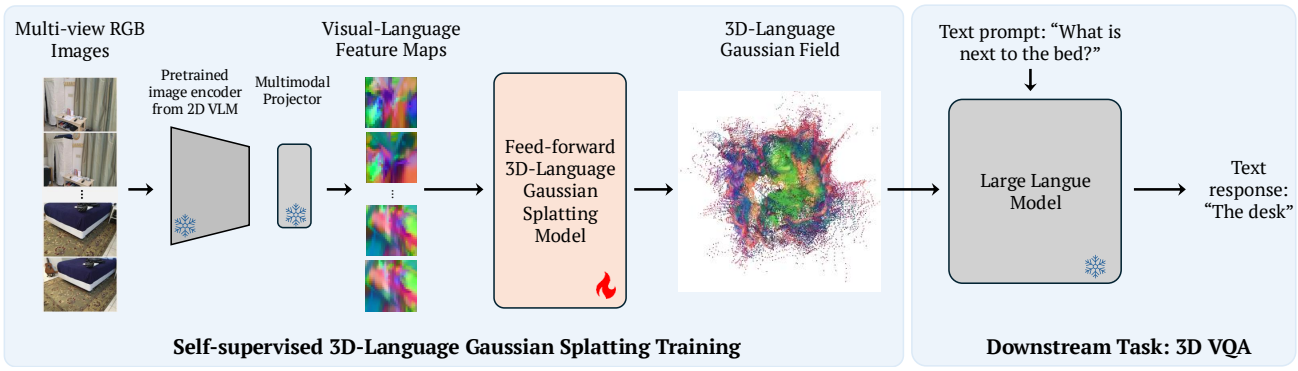

Figure 1. **SplatTalk.** We propose a self-supervised 3D-Language Gaussian Splatting model trained from multi-view RGB images. First, images are encoded using a pretrained 2D Vision-Language Model (VLM) and projected into visual-language feature maps via a multimodal projector. These feature maps are then learned within a feed-forward 3D-Language Gaussian Splatting model, producing a 3D-Language Gaussian Field that encodes spatial and semantic information in 3D space. During inference, the 3D Gaussian features are directly queried by a Large Language Model (LLM) to perform 3D question-answering (3D VQA) tasks.

Abstract

Language-guided 3D scene understanding is important for advancing applications in robotics, AR/VR, and human-computer interaction, enabling models to comprehend and interact with 3D environments through natural language. While 2D vision-language models (VLMs) have achieved remarkable success in 2D VQA tasks, progress in the 3D domain has been significantly slower due to the complexity of 3D data and the high cost of manual annotations. In this work, we introduce SplatTalk, a novel method that uses a generalizable 3D Gaussian Splatting (3DGS) framework to produce 3D tokens suitable for direct input into a pre-trained LLM, enabling effective zero-shot 3D visual question answering (3D VQA) for scenes with only posed images. During experiments on multiple benchmarks, our approach outperforms both 3D models trained specifically for the task and previous 2D-LMM-based models utilizing only images (our setting), while achieving competitive performance with state-of-the-art 3D LMMs that additionally utilize 3D inputs. Project website: https://splat-talk.github.io/

1. Introduction

The rise of Large Multimodal Models (LMMs), following the success of Large Language Models (LLMs), has become increasingly significant in a world where visual understanding is crucial. While 2D LMMs have seen rapid improvements, enabling models to analyze images and answer complex questions [21, 28, 37, 40], progress in 3D LMMs has been slower. This gap is largely due to the scarcity of 3D data compared to 2D datasets. Further, the availability of language-annotated 3D data is even more limited, as annotating 3D scenes since it is significantly more complex and expensive than labeling 2D images. These challenges have hindered the development of language-guided 3D reasoning, making 3D LMMs an important yet underexplored area in multimodal research. In this work, we focus on 3D Visual Question Answering (3D VQA) as it provides a strong benchmark for evaluating a model's ability to understand and reason about 3D scenes.

Recently, significant progress has been made in 3D multimodal learning, particularly for indoor scenes [15, 17, 27, 48]. Several approaches rely on posed images and 3D inputs, like meshes [41], point clouds [9, 17, 34] and voxels [13], to associate 3D geometry with language features. As such,

they have limited applicability in scenarios where a full 3D surface reconstruction is not available.

Previous works have explored integrating semantic features into NeRF [19] or 3D Gaussian Splatting (3DGS) [35], which rely only upon RGB image inputs. However, research on effectively associating 3D geometry with language fetures remains limited. Some methods align individual 3D locations with language features via implicit functions optimized individually for every scene [10, 19]. Others employ SAM [20] to first segment objects and object parts, followed by CLIP [36] to extract object-centric features from the generated masks. These approaches allow models to query information about individual points and/or objects. However, it disregards spatial relationships between objects, making it unsuitable for 3D VQA spatial reasoning tasks.

More recent work [21] has demonstrated the power of LLMs to answer VQA questions using only a sampling of tokens extracted from images, without any underlying 3D representation. While this approach is inspiring, it relies upon the LLM to make sense of a set of tokens that are arranged according to images rather than elements in the 3D scene. As a result, it struggles to answer queries about specific 3D locations (e.g., "What is here?") and questions requiring a holistic understanding of the entire 3D scene, where the answer depends on establishing correspondences between objects across multiple images (e.g., "What is opposite the door at the other end of the room"?). In this work, we aim to address these issues by consolidating tokens in 3D before providing them to the LLM.

Our approach is called *SplatTalk*. Given a set of posed RGB images, we integrate language features extracted from the RGB images to build a 3D Gaussian Splatting representation, from which 3D tokens can be extracted for direct input into a pretrained LLM to answer challenging 3D VQA queries (see Fig. 1). The main research contributions are the new methods for 1) integrating language features into the Gaussian representation, and 2) extracting tokens that can be provided directly to an LLM. These methods avoid the issues of concurrent work, ChatSplat [6], which produces tokens that are not directly compatible with an LLM, making language-based reasoning more difficult.

Overall, SplatTalk has a unique combination of three advantages. First, it does not require any 3D surface representation as input. Second, it leverages the full power of a pre-trained LLM to answer difficult 3D VQA questions. Third, it produces a concise 3D representation that can be queried and/or edited in 3D (e.g., cropped to a 3D region, rendered from a novel view, etc.). During experiments with 3D VQA benchmarks, we find that it outperforms 3D models trained only for the benchmark tasks, demonstrating the value of utilizing pre-trained LLMs. SplatTalk further outperforms prior image-only methods using LLMs, showing the value of building a 3D scene representation. Finally, it even achieves competitive performance with SOTA 3D LMMs that utilize 3D scene representation inputs.

In summary, our key contributions are:

- We introduce the first self-supervised 3D Gaussian-based method for large-scale zero-shot 3D VQA, SplatTalk, requiring only multi-view images without explicit depth, point clouds or 3D-language annotated supervision.
- We show that mean feature extraction from 3D Gaussians encodes holistic scene concepts for 3D VQA, supported by theoretical justification.
- We introduce entropy-adaptive token sampling, leading to better 3D VQA performance without additional training.
- We demonstrate that SplatTalk outperforms 2D LMMs and achieves performance on par with 3D models using point clouds on various 3D VQA benchmarks, demonstrating the effectiveness of our 3D-aware representations.

2. Related Work

2D LMMs. With advancements in large language models (LLMs), the ability of models to comprehend images and respond to language prompts related to their content has improved significantly. Notably, SOTA models [1, 11, 21, 30, 32] have demonstrated remarkable capabilities in processing multiple images of a 3D scene and answering questions about the shared content across these images. Although multiple RGB images inherently contain 3D information, they do not explicitly model 3D understanding, potentially leading to sub-optimal performance on tasks that require explicit 3D reasoning. In this work, we build upon the strong performance of 2D vision-language models (2D-VLMs), particularly LLaVA-OneVision [21], and integrate them with the 3D priors provided by the 3DGS framework to improve 3D spatial understanding capabilities.

3D LMMs. Recent advancements in 3D indoor scene understanding have explored integrating 3D priors into LLMs using inputs such as 3D point clouds [7, 17] or multi-view RGB images combined with depth or 3D point clouds [48]. These approaches employ various strategies to embed 3D structural information and language features into a shared representation space. While most methods [17, 48] fine-tune pretrained visual encoders (both 2D and 3D) alongside LoRA parameters or even the full set of LLM parameters, others [18, 24] opt for a more lightweight approach, training only a projection layer to map visual features into the language feature space. In contrast, our method does not require any explicit depth or 3D inputs. Instead, we leverage the 3DGS framework, which has shown strong capabilities in representing 3D scenes using only multi-view RGB images.

Generalizable 3D Gaussian Splatting. This body of work explores generalizable 3DGS as an alternative to the per-scene optimization paradigm used in traditional 3DGS methods, aiming to improve novel view synthesis across unseen scenes. Our method builds on FreeSplat [43], which is

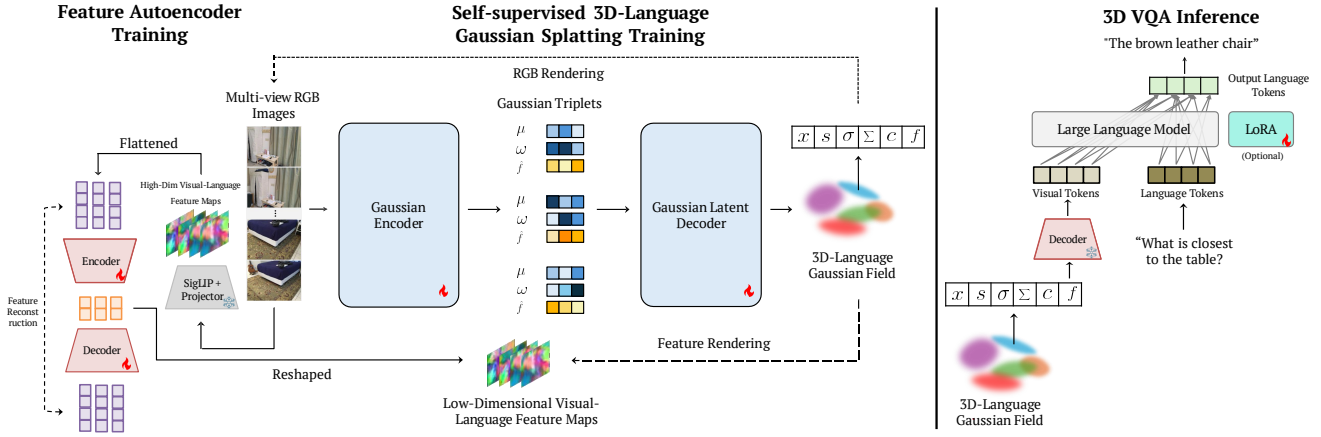

Figure 2. **Overview.** *Left*: During the self-supervised 3D-language Gaussian Splatting training phase, multiple RGB input views are first encoded into Gaussian latent features (Gaussian triplets). These latent features are then decoded into Gaussian parameters for rendering, along with a low-dimensional visual-language feature f. To ensure proper supervision of this low-dimensional feature, we train an autoencoder that maps the high-dimensional, unbounded features obtained from LLaVA-OV, specifically, the visual tokens serving as direct inputs to the LLM, onto a low-dimensional hypersphere space. *Right*: During 3D VQA inference, visual-language features are directly extracted from the 3D Gaussians. These features are then mapped back to the original high-dimensional space using the pretrained decoder and subsequently used as direct visual token inputs to the LLM. LoRA fine-tuning of the LLM is optional.

designed for sparse-view synthesis in indoor scenes. The method begins by extracting multi-scale features from input RGB images using CNN backbones. It then constructs adaptive cost volumes from neighboring views to predict depth maps. These predicted depth maps are used to back-project 2D CNN-encoded features into 3D Gaussian triplets $(\boldsymbol{\mu}, \boldsymbol{\omega}, \boldsymbol{f})$ where $\boldsymbol{\mu}$ denotes the Gaussian's center, $\boldsymbol{\omega}$ the weight, and $\boldsymbol{f}$ the latent feature. To integrate multi-view information, FreeSplat introduces a Pixel-wise Triplet Fusion (PTF) module that uses a lightweight GRU to align and fuse overlapping 3D Gaussians across views efficiently. Finally, the fused triplets are passed through an MLP to decode rendering parameters, $\boldsymbol{\Sigma}, \boldsymbol{\alpha}, \boldsymbol{s}$. Our approach extends the FreeSplat framework to enable the learning of generalizable 3D Gaussian language fields.

Similar to FreeSplat [43], [5, 8] leverage multi-view stereo (MVS) cues to refine depth estimation and improve geometric consistency in 3D Gaussian representations. These methods primarily focus on RGB novel view synthesis from sparse inputs (2–10 views). Our work builds on the SOTA model, FreeSplat [43], but diverges significantly in both objectives and methodology. First, while previous methods focus on photorealistic RGB reconstruction, we aim at semantic 3D scene understanding. Second, instead of reconstructing only a localized portion of the scene as shown in these works, our goal is to represent entire 3D environments, capturing both spatial and semantic relationships holistically.

Language-guided 3D Gaussian Splatting. LangSplat [35] is one of the first approaches that integrates semantic features into 3DGS. However, this method requires training on individual scenes, which is significantly expensive, and lacks generalization to unseen scenes. Similarly, follow-up works [23, 45] explore 3DGS for open-vocabulary object segmentation but do not focus on reasoning tasks like 3D VQA. Further, they primarily rely on scene-specific training, limiting their scalability. In contrast, we propose a fully self-supervised framework that learns 3D-aware visual representations without requiring per-scene training. While some works [12, 42] explores a generalizable 3DGS and semantic modeling approach, they primarily focus on open-vocabulary segmentation, without addressing the broader challenges of free-form language reasoning required for 3D VQA.

ChatSplat [6] is a concurrent work that explores natural language conversation within the 3DGS framework. However, it optimizes on a per-scene basis and requires rendering 3D features into 2D feature maps during inference. Additionally, its training pipeline is not fully end-to-end, as it first trains the RGB Gaussians separately before incorporating language features. In contrast, our approach leverages scene latent features that encapsulate both RGB and semantic information, enabling the joint optimization of both modalities in a fully end-to-end manner. Our approach extends the FreeSplat framework to enable the learning of generalizable 3D Gaussian language fields.

3. Method

The problem we address is formulated as follows: Given a sequence of N posed RGB images captured from a 3D indoor scene, along with language prompts related to the scene's content, the goal is to generate the corresponding language responses. That is, we address 3D VQA tasks with solely posed RGB images as inputs.

We provide an overview of our approach in Fig. 2, which

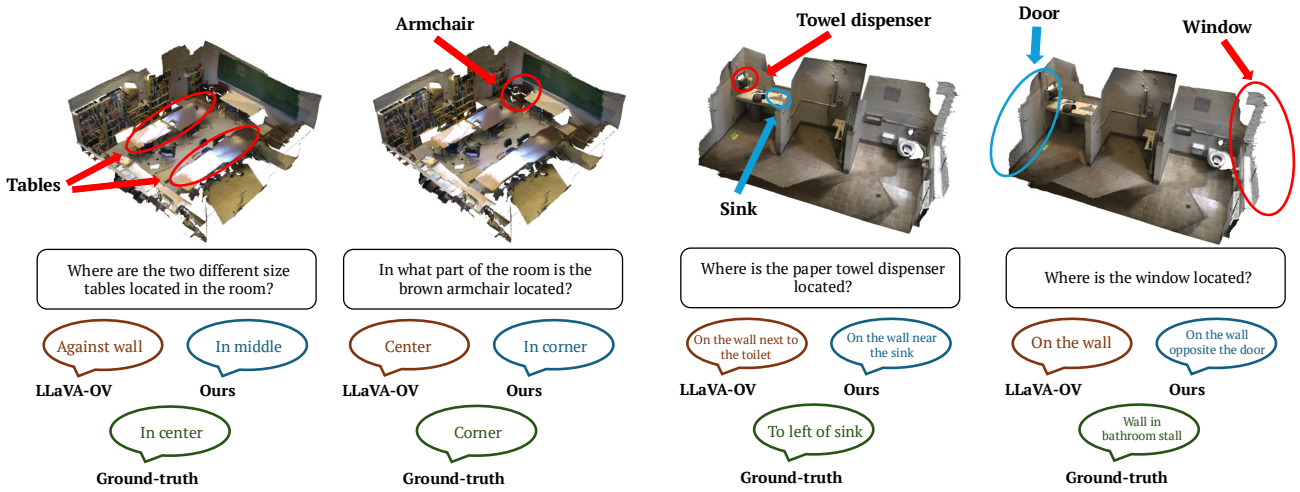

Figure 3. **Qualitative results on ScanQA**, scene0030_00 and scene0084_00. We compare responses from LLaVA-OV, our model (Ours), and the ground truth (GT) for spatial reasoning questions in 3D VQA. Each scene highlights the referenced objects with red circles, and key relative objects are marked in blue. The answers from each model are displayed in color-coded speech bubbles: LLaVA-OV (brown), Ours (blue), and GT (green). With its 3D-aware representation, our model exhibits improved spatial reasoning capabilities, accurately identifying relationships between objects that are far apart and may never co-occur in the same image (e.g., door and window). More examples in Supp.

consists of three main steps: 1) Training the feature autoencoder, 2) Self-supervised 3D-Language Gaussian Splatting Training and 3) 3D VQA Inference. First, we extract the high-dimensional language features from LLaVA-OV [21] for each RGB image by using the pretrained image encoder and multimodal projector. Since these features are high-dimensional, unbounded and sparse, we train an autoencoder to project them into a more compact, low-dimensional space. These low-dimensional features then serve as pseudo ground-truth for training the language representations in SplatTalk. SplatTalk architecture is a feed-forward network consisting of a Gaussian encoder and decoder that predicts Gaussian parameters for rendering. During 3D VQA inference stage, we extract the trained language features from the 3D Gaussians and directly feed them into the LLM, along with the language prompt, to generate free-form responses.

3.1. Integrating Language Features

An important step of our computational pipeline is the self-supervised method to integrate language features into a generalizable 3DGS framework. We identify four key factors that shape our approach for this step: (1) Leveraging features that align with the LLM's input space for direct language reasoning, (2) Efficient feature dimensionality reduction for effective CUDA-based rendering, and (3) Joint training of RGB and semantic features for coherent multimodal representation learning.

Visual Tokens for 2D Pseudo Ground-truth Features. We extract the visual tokens used as inputs to the LLM for training our 3DGS framework. Our empirical findings indicate that extracting features directly from the visual encoder for training 3DGS requires retraining the multimodal projector layer which aligns the visual with language spaces. This observation is further corroborated by [6], which demonstrates that features extracted before the multimodal projector cannot be seamlessly integrated into the language model. A possible explanation is that the image encoder's native feature space (before the projector) is high-dimensional and unbounded, leading to sparser features with greater variance across different dimensions. As a result, the projector becomes more sensitive to subtle variations in the input features, making it difficult to directly map them into a token representation that aligns with the LLM's understanding. In contrast, the visual tokens generated after the projector are already well-aligned with the language model's latent space, allowing the LLM to interpret and reason with them directly.

Feature Dimensionality Reduction. Building on the previous factor, directly optimizing these language embeddings within a differentiable rendering pipeline is challenging because they are extremely high-dimensional (e.g., 3584 dimensions), sparse, and contain unbounded values, making learning both unstable and impractical. Therefore, we learn a generalizable autoencoder that projects these sparse, high-dimensional visual tokens into a compact unit hypersphere with much lower dimension (e.g., 256 dimensions), ensuring feature smoothness. Note that the language representations from LLaVA, which support reasoning and free-form language comprehension, are significantly more complex than those from CLIP or segmentation-based models like SAM. Different from prior works that compress semantic features into extremely low-dimensional vectors (e.g., 3–16 dimensions [35, 45]), which simplifies rendering but results in significant information loss, our approach prior-

Table 1. **Evaluation of 3D VQA on ScanQA [2] and SQA3D [31] datasets.** We compare SplatTalk against various baselines, including 3D LMMs, which take point clouds (PC) or both point clouds and images (PC+I) as visual inputs, and 2D LMM-based models, which rely solely on images (I). As a 2D LMM-based approach, SplatTalk outperforms other 2D LMM-based models, surpasses some 3D LMMs, and achieves competitive performance against the strongest 3D LMM baselines. Our models are highlighted in blue, bold indicates best performance among the 2D-LMM methods.

Method	Modality	ScanQA (val)					SQA3D (test)	
		CIDEr	METEOR	ROUGE	EM@1	EM@1-R	EM@1	EM@1-R
Specialist Models								
ScanQA [2]	PC	64.9	13.1	33.1	21.1	-	47.2	-
ScanRefer+MCan [46]	PC	55.4	11.5	30.0	18.6	-	-	-
3D-VisTA [49]	PC	69.6	13.9	35.7	22.4	-	48.5	-
SQA3D [31]	PC	-	-	-	-	-	46.6	-
Generalist 3D LMMs								
Chat-Scene [16]	PC+I	87.7	18.0	41.6	21.6	-	54.6	57.5
Scene-LLM [13]	PC+I	80.0	15.8	35.9	-	-	53.6	-
Grounded 3D-LLM [9]	PC	72.7	-	-	-	-	-	-
Chat-3D-v2 [44]	PC	77.1	16.1	40.1	-	-	54.7	-
Finetuned 3D LMMs								
3D-LLM [15]	PC+I	69.4	14.5	35.7	20.5	-	-	-
LEO [17]	PC+I	101.4	20.0	49.2	24.5	47.6	50.0	52.4
Scene-LLM [13]	PC+I	80.0	16.6	40.0	27.2	-	54.2	-
LL3DA [7]	PC	76.8	15.9	37.3	-	-	-	-
Chat-3D [44]	PC	53.2	11.9	28.5	-	-	-	-
2D LMM-Based								
VideoChat2 [25]	I	49.2	9.5	28.2	19.2	-	37.3	-
GPT-4V [32]	I	59.6	13.5	33.4	-	-	-	-
Claude [1]	I	57.7	10.0	29.3	-	-	-	-
LLaVA-NeXT-Video [22]	I	46.2	9.1	27.8	-	-	34.2	-
LLaVA-OV [21]	I	50.0	13.0	29.4	15.6	32.5	25.4	35.8
SplatTalk	I	61.7	14.2	32.7	17.1	32.2	26.1	41.3
SplatTalk-ScanQA-FT	I	77.1	15.4	**38.7**	22.3	**38.3**	42.0	44.7
SplatTalk-3DVQA-FT	I	**77.5**	**15.6**	38.5	**22.4**	**38.3**	**47.6**	**49.4**

itizes preserving rich language features by maintaining a higher-dimensional representation (e.g., 256). To efficiently render high-dimensional feature maps on CUDA, we draw inspiration from [47] and implement a parallel CUDA rasterizer pipeline, where semantic features and RGB are rendered simultaneously, using shared Gaussian parameters.

Unlike previous methods [6, 35] that train a separate autoencoder for each individual scene, we train a single autoencoder across all training scenes, each consisting of multiple frames. This approach improves the autoencoder's generalizability, reducing computational overhead by eliminating the need for retraining to compress features from unseen scenes. Training the autoencoder separately from the 3DGS model ensures that it is dedicated solely to compressing high-dimensional feature vectors while preserving the essential scene information without being influenced by the complexities of the rendering pipeline. Our empirical findings further indicate that jointly training a feature decoder to project low-dimensional embeddings back to the high-dimensional space introduces significant instabilities in the training process.

Joint-training RGB and Language. In contrast to previous approaches that first train RGB Gaussians and then freeze them before integrating language features [6, 35], we adopt a joint training strategy, where RGB and language features are learned simultaneously. This is due to the fact that visual RGB and semantic features play a complementary role in effectively representing 3D scenes [33]. In feedforward models like FreeSplat, where the scene is first encoded into a Gaussian latent space before being decoded into specific Gaussian parameters for RGB rendering, these Gaussian latent features inherently capture holistic scene information, including both appearance (RGB) and semantics. Based on this observation, we introduce a new Gaussian parameter decoder head that predicts semantic features for each 3D Gaussian alongside rendering parameters, enabling joint op-

Table 2. **Evaluation of 3D VQA on MSR3D [27].** We compare SplatTalk against various baselines, all evaluated in a zero-shot setting. T denotes models that rely solely on text inputs without interleaving images, PC indicates models that use point cloud as the visual input, and I represents models that process only image-based inputs. Our approach, highlighted in blue, outperforms all baselines.

Model	Input	Scene	Setting	Counting	Existence	Attributes	Spatial	Navigation	Others	Overall
LEO [17]	T	PC	zero-shot	0.80	15.5	11.8	7.3	2.3	15.3	7.8
LLaVA-OV [21]	T	I	zero-shot	18.6	31.2	24.8	19.5	16.7	36.3	24.0
SplatTalk	T	I	zero-shot	19.6	60.3	**44.0**	**35.8**	**35.5**	**61.8**	**41.8**
SplatTalk-ScanQA-FT	T	I	zero-shot	**28.9**	**66.0**	43.2	28.3	33.8	60.0	41.5
SplatTalk-3DVQA-FT	T	I	zero-shot	24.3	57.0	36.1	7.8	16.4	35.0	26.5

timization. Unlike approaches that explicitly model these interactions [33, 39] using extra cross-attention or projection layers, SplatTalk implicitly integrates semantic and visual information within the Gaussian representation. This paradigm naturally facilitates multimodal interactions between RGB and language features.

To incorporate more views and comprehensively cover the scene, we reduce the input resolution of the RGB images. This adjustment aligns well with the inherently low spatial resolution of the visual tokens extracted from LLaVA-OV, and also prevents the model from disproportionately allocating resources to modeling RGB features, thus encouraging a balanced representation of semantic and RGB information.

These design choices establish our approach as a fundamentally novel method for integrating language features into 3DGS, grounded in first-principles observations rather than heuristic adaptations. By addressing the core challenges of multimodal alignment, feature compression, and joint optimization, our approach is not merely a trivial extension of existing works, but a carefully designed framework that ensures both effectiveness and stability in learning rich, spatially-aware semantic representations.

3.2. Extracting 3D Language Features

During inference, we directly extract language features at the mean position of each 3D Gaussian. Although the rendering process incorporates each Gaussian's covariance and opacity when projecting 3D features to 2D, we demonstrate that *scene semantics can be effectively captured by the mean feature of each Gaussian, $\boldsymbol{f}_i$, alone, without relying on these additional factors.* This is supported by the conceptual similarity between our training optimization process and the Expectation-Maximization (EM) algorithm. For simplification, let $\mathcal{L} = \sum_{t=1}^{T} \|\boldsymbol{F}_t(x) - \boldsymbol{F}_t^{gt}(x)\|^2$ be the loss, where $\boldsymbol{F}_t(x)$ is the predicted rendered feature map of frame t at pixel x and $\boldsymbol{F}^{gt}$ is the pseudo ground-truth feature map obtained from LLaVA-OV. Replacing $\boldsymbol{F}_t(x)$ by the rendering equation and solving for the optimal $\boldsymbol{f}_i^*$, we obtain $\boldsymbol{f}_i^* = \frac{\sum_{t,x} R_i(t,x) \boldsymbol{F}_t^{gt}(x)}{\sum_{t,x} R_i(t,x)}$ where $R_i(t,x) = \frac{\alpha_i \mathcal{N}(x;\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)}{\sum_j \alpha_j \mathcal{N}(x;\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}$ is the Gaussian's contribution to pixel x at viewpoint t. This mirrors the EM procedure: estimating R by rendering (E-step) and updating f_i by minimizing reconstruction loss (M-step). Consequently, the optimal 3D feature $\boldsymbol{f}_i^*$ for each Gaussian is a weighted sum of the 2D feature maps, and the set of 3D Gaussian features $\{\boldsymbol{f}_i\}_{i=1}^N$ encodes the scene's semantic holistically from T 2D inputs. These features are then projected back into the original language embedding space using the pretrained decoder and directly provided as inputs to the LLM for downstream 3D-language tasks.

In open-vocabulary segmentation, querying Gaussian centers is effective [4, 35, 45] because each Gaussian is assumed to represent a distinct, meaningful object or region. These methods use object-centric embeddings (e.g., SAM-masked CLIP or LSeg) and retrieve objects via cosine similarity to text queries. In contrast, 3D reasoning tasks lack this one-to-one correspondence. Each Gaussian encodes parts of a scene-level representation, where meaning stems from spatial and relational structure. Querying single Gaussians is thus insufficient for context-dependent reasoning. While our method also queries at Gaussian means, this is not trivial; our EM-based formulation offers a principled justification.

3.3. 3D VQA Inference

Entropy Adaptive Gaussian Sampling. SplatTalk leverages triplet fusion adopted in FreeSplat [43], which reduces the number of Gaussians. However, complex scenes still require significantly more Gaussians than the LLM's visual token limit, and not all are equally informative. To address this, we propose entropy-based sampling: selecting the top k Gaussians with the highest language feature entropy, where k matches the LLM's visual token capacity.

3D Tokens as Visual Inputs. After 3DGS training, each 3D Gaussian encodes a feature that is a weighted aggregation of 2D semantic information across all input views (Sec. 3.2). These features are also spatially grounded in their 3D coordinates. Because each token corresponds to a specific location in 3D space, its position is already implicitly encoded in the feature through the rendering and fusion process. Unlike 2D patches in a fixed grid, these 3D tokens form an unordered set, similar to a point cloud where spatial relationships are captured by the features themselves, not by sequence position. Therefore, we do not require explicit positional encodings or a fixed input order. The 3D tokens,

extracted by sampling features at each Gaussian's mean, are directly fed into the LLM.

3.4. Optional - Finetuning on 3D VQA Datasets

While SplatTalk is capable of performing the 3D VQA task in a zero-shot manner, it also allows for flexible fine-tuning on 3D VQA datasets to enhance alignment and reasoning capabilities. Rather than training the model end-to-end, which is both computationally expensive and prone to overfitting on limited datasets, we optimize only the LoRA parameters applied to the LLM during fine-tuning.

4. Experiments

4.1. Implementation Details

We train SplatTalk with 500 scenes from the ScanQA training set. The initial self-supervised stage focuses purely on learning the 3D scene representation, without leveraging any text-annotated information. For each scene, we uniformly sample 100 views around the scene and extract their per-patch features by passing them through the visual encoder of LLaVA-OV followed by the multi-modal projector. The subsequent SplatTalk model is trained on a single NVIDIA H100 (80GB) GPU. Both the LoRA fine-tuning and inference are also conducted on a single H100. More details can be found in Supp. Material.

4.2. Datasets and Metrics

Evaluation Dataset. We evaluate our method on ScanNet-based QA datasets: ScanQA [2], SQA3D [31], and MSR3D [27]. ScanQA is a benchmark designed for question answering in 3D indoor scenes, containing detailed spatial and semantic information extracted from ScanNet. SQA3D and MSR3D further extend this by introducing scenarios that describe an embodied agent's 3D position and state.

Metrics. For ScanQA and SQA3D, we evaluate performance using standard n-gram-based metrics, specifically CIDEr [38], METEOR [3], ROUGE [26], EM@1 [14], and EM@1-Refined for ScanQA, and EM@1 and EM@1-Refined for SQA3D. In contrast, for MSR3D, we utilize the LLM-Match score [29] as recommended by the authors.

4.3. Performance on 3D Question Answering Tasks

ScanQA. We evaluate SplatTalk's performance against several model families, each representing different levels of specialization and generalization. "Specialist Models" refer to models specifically trained to solve ScanQA without exposure to other tasks. "Generalist 3D LMMs" consist of models trained across a diverse range of 3D vision-language tasks, enabling broader generalization. "Finetuned 3D LMMs" are models that have been further fine-tuned on the ScanQA training set to optimize performance for this particular dataset. Lastly, "2D LMM-Based Models" include models that process only image inputs without leveraging depth or explicit 3D information. SplatTalk and its fine-tuned variants fall within this category.

We evaluate three different variants of our model on this task. The base model (SplatTalk) is trained solely to encode the semantic representation of the scene from multiple RGB images using 3DGS, without exposure to question-answering tasks. Note that this model is trained entirely in a self-supervised manner, without any access to language supervision or textual annotations during training. To further adapt it to downstream 3D VQA tasks, we fine-tune this base model on the ScanQA training set, referring to it as SplatTalk-ScanQA-FT, and on the combined training sets of ScanQA and SQA3D, denoting as SplatTalk-3DVQA-FT. We present results in Table 1.

Compared to LLaVA-OV, whose 2D language feature maps are used to train SplatTalk, our base model, which benefits from a more 3D-aware representation due to the 3D Gaussian training process, achieves significantly higher performance (e.g. 61.7 vs 50.0 on CIDEr). Our language fine-tuned model variants, SplatTalk-ScanQA-FT and SplatTalk-3DVQA-FT, improve performance further. Notably, despite not being trained with depth supervision or explicit 3D information, our models achieve comparable results to both Generalist 3D LMMs and Finetuned 3D LMMs, while outperforming Specialist Models. Additionally, among 2D LMM-Based approaches, our base model without being fine-tuned with 3D VQA datasets outperforms other methods, with significantly higher performance on some metrics, including both open-source and closed-source models. We show qualitative results in Fig. 3.

SQA3D. Similar to our evaluation on ScanQA, we compare our model against both 3D LMMs and 2D LMMs. Compared to 2D LLaVA-OV features, our base model with 3D-aware features achieves a significant performance gain, highlighting the advantage of incorporating 3D information via 3DGS which aligns with our findings on ScanQA. Further fine-tuning on ScanQA alone and on the combined ScanQA + SQA3D dataset leads to substantial improvements. Notably, SplatTalk-3DVQA-FT surpasses specialist models that leverage 3D point cloud inputs, demonstrating the strength of our approach in 3D scene understanding.

MSR3D. We evaluated our model's performance against LLaVA-OV and LEO[1] [17] on the Scannet split of MSR3D. Importantly, all comparisons are conducted in a zero-shot setting, as none of the models have been trained on MSR3D. This is an interleaved dataset where inputs consist of both text and images. However, since all models are tested in a zero-shot manner, they rely solely on text inputs, where the agent's state descriptions and the question are combined into

[1]The reported number reflects performance across all splits of MSR3D. However, since LEO was trained on ScanNet and 3RScan, its performance remains largely consistent across splits.

Table 3. **Ablation Study on Gaussian Sampling Strategies.** Our entropy-based sampling strategy consistently outperforms across multiple metrics.

Sampling Method	EM@1	METEOR	ROUGE	SPICE	Avg
Random	17.0	13.8	**62.1**	14.0	26.73
Point Density	16.2	14.0	59.7	14.1	26.00
FPS	**17.7**	13.8	62.0	13.6	26.78
Entropy	17.1	**14.6**	61.7	**14.2**	**26.90**

Table 4. **Ablation Study on Visual Input Length.** We compare model performance using 729 tokens (equivalent to a single image) versus $32,076$ tokens (corresponding to 44 images), analyzing the impact of visual context on reasoning.

Visual Length	ScanQA (val)		SQA3D (test)		MSR3D (test)	
	EM@1	EM@1-R	EM@1	EM@1-R	EM@1	EM@1-R
729 tokens	16.2	30.7	25.8	38.0	7.9	17.7
$32,076$ tokens	**17.1**	**32.2**	**26.1**	**41.3**	**14.1**	**23.3**

a single text prompt. On the visual processing side, LEO operates on 3D point cloud inputs, whereas both LLaVA-OV and our model rely only on RGB images.

The results, presented in Table 2, demonstrate that our base model outperforms LLaVA-OV across all question types. Notably, aside from the "counting" category, our base model achieves approximately twice the performance of LLaVA-OV on other question types. Both models significantly outperform LEO, which struggles due to the absence of fine-tuning on this dataset.

Interestingly, we observe a slight performance drop when comparing SplatTalk with SplatTalk-ScanQA-FT. However, this drop becomes much more pronounced when compared with SplatTalk-3DVQA-FT. A plausible explanation is that fine-tuning on other QA datasets introduces overfitting, causing the model to adapt too specifically to those datasets. Furthermore, the observed performance drop suggests that MSR3D differs significantly from ScanQA and SQA3D, making direct knowledge transfer less effective. This distinction may also explain why LEO struggles to generalize in a zero-shot setting. Despite being trained on multiple QA datasets, none of them closely resemble MSR3D, limiting its ability to adapt without fine-tuning.

4.4. Ablation Studies

Gaussian Sampling Strategies. We explore various Gaussian sampling strategies for selecting tokens as inputs to the LLM, including: (1) Random Sampling—Gaussians are selected randomly across the scene. (2) Point Density Adaptive Sampling—Gaussians are sampled proportionally to the density of 3D points. (3) Furthest Point Sampling (FPS)—Gaussians are iteratively selected at locations maximally distant from previously chosen points. (4) Entropy Adaptive Sampling—Gaussians are prioritized based on language feature entropy, selecting points with the highest uncertainty.

We conduct experiments on the ScanQA validation set. The results, presented in Table 3, demonstrate that entropy-adaptive sampling outperforms other sampling strategies, followed by FPS. This finding highlights entropy-adaptive sampling as the most effective approach for selecting informative tokens, making it the optimal strategy for downstream 3D VQA tasks.

Length of Visual Tokens. We evaluate the performance of SplatTalk on ScanQA, SQA3D, and MSR3D under two different settings: (1) using only 1 image-worth of visual tokens (729 tokens) and (2) using the full context length corresponding to 44 images-worth of visual tokens ($32,076$ tokens). The results, presented in Table 4, evaluate model performance using EM@1 and EM@1-Refined, as these metrics offer a reliable measure of how precisely model predictions align with ground truth answers, both in literal accuracy and permissible variations in phrasing. We employ entropy-adaptive sampling (see Sec. 3.3), which prioritizes high-information tokens and remains deterministic given a fixed set of visual tokens.

A key observation is that performance improved across all datasets as the number of input tokens increased, with MSR3D showing the largest gain, nearly doubling in EM@1. In contrast, ScanQA and SQA3D saw only marginal improvements, even with significantly more tokens. This suggests that additional scene context may not substantially improve performance, possibly due to: (1) Dataset limitations where the tasks may not require rich spatial context; or (2) Model limitations where the model may be unable to effectively leverage extra visual tokens for spatial reasoning.

5. Conclusion

In this work, we introduce the first method to integrate free-form language features into a 3D Gaussian Splatting framework, enabling zero-shot generalization for 3D Visual Question Answering (3D VQA) without requiring 3D point clouds or depth maps. Our approach surpasses methods that rely solely on 2D language feature maps and achieves performance on par with 3D LMMs that explicitly utilize 3D structural information across multiple 3D VQA benchmarks. We hope this work inspires further research into leveraging multi-view RGB inputs to encapsulate rich semantic representations of 3D scenes, ultimately enhancing 3D VQA capabilities without the reliance on explicit geometric priors.

Acknowledgment

We thank the creators of the datasets and codebases used in our work, ScanQA, SQA3D, MSR3D, ScanNet, LLaVA-OV, and FreeSplat, for making their resources publicly available. We are also grateful to the anonymous reviewers for their constructive feedback during the review process. Special thanks to Xiang Li and James M. Rehg for the insightful discussions.

References

[1] Anthropic. Claude 3. https://www.anthropic.com/index/claude, 2024. Accessed: 2024-02-28. 2, 5

[2] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139, 2022. 5, 7

[3] Satanjeev Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, 2005. 7

[4] Yash Bhalgat, Iro Laina, João F Henriques, Andrew Zisserman, and Andrea Vedaldi. N2f2: Hierarchical scene understanding with nested neural feature fields. In *European Conference on Computer Vision*, pages 197–214. Springer, 2024. 6

[5] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19457–19467, 2024. 3

[6] Hanlin Chen, Fangyin Wei, and Gim Hee Lee. Chatsplat: 3d conversational gaussian splatting. *arXiv preprint arXiv:2412.00734*, 2024. 2, 3, 4, 5

[7] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26428–26438, 2024. 2, 5

[8] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pages 370–386. Springer, 2024. 3

[9] Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Runsen Xu, Ruiyuan Lyu, Dahua Lin, and Jiangmiao Pang. Grounded 3d-llm with referent tokens. *arXiv preprint arXiv:2405.10370*, 2024. 1, 5

[10] Francis Engelmann, Fabian Manhardt, Michael Niemeyer, Keisuke Tateno, Marc Pollefeys, and Federico Tombari. Opennerf: open set 3d neural scene segmentation with pixel-wise features and rendered novel views. *arXiv preprint arXiv:2404.03650*, 2024. 2

[11] Gemini Team et al. Gemini: A family of highly capable multimodal models, 2024. 2

[12] Zhiwen Fan, Jian Zhang, Wenyan Cong, Peihao Wang, Renjie Li, Kairun Wen, Shijie Zhou, Achuta Kadambi, Zhangyang Wang, Danfei Xu, et al. Large spatial model: End-to-end unposed images to semantic 3d. *Advances in Neural Information Processing Systems*, 37:40212–40229, 2025. 3

[13] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024. 1, 5

[14] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6904–6913, 2017. 7

[15] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023. 1, 5

[16] Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. *arXiv preprint arXiv:2312.08168*, 2023. 5

[17] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023. 1, 2, 5, 6, 7

[18] Weitai Kang, Haifeng Huang, Yuzhang Shang, Mubarak Shah, and Yan Yan. Robin3d: Improving 3d large language model via robust instruction tuning, 2025. 2

[19] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739, 2023. 2

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 2

[21] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 2, 4, 5, 6

[22] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024. 5

[23] Haijie Li, Yanmin Wu, Jiarui Meng, Qiankun Gao, Zhiyao Zhang, Ronggang Wang, and Jian Zhang. Instancegaussian: Appearance-semantic joint gaussian representation for 3d instance-level perception. *arXiv preprint arXiv:2411.19235*, 2024. 3

[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. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2

[25] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 5

[26] Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, pages 74–81, 2004. 7

[27] Xiongkun Linghu, Jiangyong Huang, Xuesong Niu, Xiaojian Shawn Ma, Baoxiong Jia, and Siyuan Huang. Multi-modal situated reasoning in 3d scenes. *Advances in Neural Information Processing Systems*, 37:140903–140936, 2025. 1, 6, 7

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1

[29] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. Accepted as Oral Presentation. 7

[30] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024. 2

[31] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022. 5, 7

[32] OpenAI. Gpt-4 with vision (gpt-4v). https://openai.com/research/gpt-4, 2023. Accessed: 2024-02-28. 2, 5

[33] Qucheng Peng, Benjamin Planche, Zhongpai Gao, Meng Zheng, Anwesa Choudhuri, Terrence Chen, Chen Chen, and Ziyan Wu. 3d vision-language gaussian splatting. *arXiv preprint arXiv:2410.07577*, 2024. 5, 6

[34] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 815–824, 2023. 1

[35] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20051–20060, 2024. 2, 3, 4, 5, 6

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[37] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024. 1

[38] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015. 7

[39] Peihao Wang, Zhiwen Fan, Zhangyang Wang, Hao Su, Ravi Ramamoorthi, et al. Lift3d: Zero-shot lifting of any 2d vision model to 3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21367–21377, 2024. 6

[40] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2025. 1

[41] Xingrui Wang, Wufei Ma, Zhuowan Li, Adam Kortylewski, and Alan L Yuille. 3d-aware visual question answering about parts, poses and occlusions. *Advances in Neural Information Processing Systems*, 36:58717–58735, 2023. 1

[42] Xingrui Wang, Cuiling Lan, Hanxin Zhu, Zhibo Chen, and Yan Lu. Gsemsplat: Generalizable semantic 3d gaussian splatting from uncalibrated image pairs, 2024. 3

[43] Yunsong Wang, Tianxin Huang, Hanlin Chen, and Gim Hee Lee. Freesplat: Generalizable 3d gaussian splatting towards free view synthesis of indoor scenes. *Advances in Neural Information Processing Systems*, 37:107326–107349, 2025. 2, 3, 6

[44] 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. *arXiv preprint arXiv:2308.08769*, 2023. 5

[45] Yanmin Wu, Jiarui Meng, Haijie Li, Chenming Wu, Yahao Shi, Xinhua Cheng, Chen Zhao, Haocheng Feng, Errui Ding, Jingdong Wang, et al. Opengaussian: Towards point-level 3d gaussian-based open vocabulary understanding. *Advances in Neural Information Processing Systems*, 37:19114–19138, 2025. 3, 4, 6

[46] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290, 2019. 5

[47] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21676–21685, 2024. 5

[48] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 2024. 1, 2

[49] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2911–2921, 2023. 5