SplatTalk: 3D VQA with Gaussian Splatting

Supplementary Material

1. Data Details

We train SplatTalk on 500 ScanNet scenes from the ScanQA training set. This stage does not involve any language supervisory signal and is purely self-supervised. We evaluate our method and baselines on 3 datasets: ScanQA [1] validation set, SQA3D [11] test set, and MSR3D [8] test set. The ScanQA validation set comprises 71 ScanNet scenes, while the SQA3D and MSR3D test sets include 67 scenes, with significant overlap with ScanQA validation set.

2. Training & Model Details

SplatTalk Training. The training of SplatTalk consists of the following main steps: 1) Extract high-dimensional language features for each RGB image using LLaVA-OV [6], 2) Train an autoencoder to compress the high-dimensional features obtained from step 1 into a more compact space, and 3) Train the 3D Gaussian Splatting (3DGS) model.

First, to extract language-informed visual features, we leverage LLaVA-OV [6], which processes RGB images through its SigLIP [14]-based vision encoder, followed by a multimodal projector. This transformation ensures that the extracted visual tokens are natively structured for the LLM's input space, allowing seamless processing and comprehension by the underlying Qwen2 [3] model. These features serve as pseudo 2D ground truth, guiding the training of the 3D-language Gaussian Splatting model.

After obtaining these high-dimensional semantic features, we train an autoencoder to compress them while preserving their most informative components. Each feature map of size $27 \times 27 \times 3584$, where 3584 represents the feature depth, is then flattened, resulting in 729 feature vectors, each of size 1×3584.

The autoencoder is composed of multiple linear layers, each interleaved with a batch normalization (BN) layer and a GeLU activation function. Specifically,

$$\begin{aligned}\text{Encoder} : (3584) &\to (2048) \to \text{BN} \to \text{GeLU} \to (1024) \\ &\to \text{BN} \to \text{GeLU} \to (512) \to \text{BN} \to \text{GeLU} \to (256)\end{aligned}$$

$$\begin{aligned}\text{Decoder} : (256) &\to (512) \to \text{GeLU} \to (1024) \to \text{GeLU} \\ &\to (2048) \to \text{GeLU} \to (2048) \to \text{GeLU} \to (3584)\end{aligned}$$

where (3584) indicates the size of the linear layer, BN stands for Batch Norm and GeLU stands for the GeLU activation function. The compressed features are normalized to reside within a unit hypersphere, which we find to be important for stabilizing the 3DGS training process.

We train the autoencoder with batch size 256, using AdamW [10] optimizer with learning rate 0.0001 for 100 epochs on a single NVIDIA H100 (80GB) GPU.

SplatTalk builds upon the FreeSplat [13] architecture, extending it with a language feature head attached to the Gaussian latent decoder. This additional head predicts a 256-dimensional feature vector for each Gaussian, alongside the standard Gaussian parameters used for RGB splatting.

For training the 3DGS component, we uniformly sample 100 views per scene. Each input image is resized to 32×32 resolution to balance efficiency and feature retention. The language features are trained using compressed pseudo 2D ground truth as supervisory signal. Inspired by [15], we implement a parallel CUDA rasterizer pipeline, allowing RGB and language features to be rendered in parallel using the shared Gaussian parameters. This approach is memory efficient and allows for rendering higher-dimensional features than the traditional rendering pipeline.

SplatTalk is trained using the Adam optimizer with an initial learning rate of $1e-4$, which decays smoothly following a cosine decay schedule. We employ a combination of photometric and semantic losses to effectively train SplatTalk. The photometric loss is computed using a combination of MSE and LPIPs, and semantic loss using MSE and cosine distance. The final loss is

$$\mathcal{L} = \|I - \hat{I}\|^2 + 0.05 \cdot LPIPS + \|F - \hat{F}\|^2 + 1 - \cos\left(F, \hat{F}\right)$$

where I and F are the ground-truth RGB image and the pseudo ground-truth language feature map respectively. $\hat{I}$ and $\hat{F}$ correspond to the predicted RGB image and language feature map respectively. We train SplatTalk on one NVIDIA H100 (80GB) GPU.

SplatTalk Inference. We begin by extracting the reconstructed 3D Gaussians that represent the entire scene. Next, we retrieve the language features corresponding to each 3D Gaussian, capturing the semantic information embedded within the scene representation.

3D VQA Inference. We sample $32,076$ tokens from the set of Gaussian tokens obtained from the SplatTalk inference step. This is equivalent to 44 image tokens, which is the context window size of the LLM. The optimal sampling method we use is entropy-based sampling. We first compute the entropy for each 3D Gaussian. Then we rank them in descending order from highest to lowest entropy. Finally, we sample the top $32,076$ tokens from this list as the input visual tokens to the LLM, along with the language tokens.

3D VQA Fine-tuning. During fine-tuning we employ LoRA [5] with $r = 16$ and $\alpha = 64$. The model is fine-tuned

for 1 epoch with learning rate $1e-5$ on a single NVIDIA H100 (80GB) GPU.

3. Evaluation Metrics

We evaluate model performance on ScanQA using n-gram-based metrics, including CIDEr [12], METEOR [2], ROUGE [7], EM@1 [4], and EM@1-Refined. For SQA3D, we report EM@1 and EM@1-Refined, while for MSR3D, we follow the authors' recommendation and use the LLM-Match metric [9] to assess answer quality.

Specifically, CIDEr evaluates the generated response by comparing n-gram similarity against multiple reference responses, assigning higher weight to words that frequently appear across multiple references. In contrast, METEOR aligns words between the generated response and reference answers using exact matches, stemming, synonyms, and paraphrasing, providing a more flexible and semantically aware evaluation. It computes both precision and recall, with a higher emphasis on recall, as human evaluation favors covering all key reference words. METEOR captures both semantic and syntactic similarity, making it more robust than purely n-gram-based metrics. ROUGE-L, on the other hand, measures similarity using the Longest Common Subsequence (LCS) between the generated response and the reference answers.

On the other hand, EM@1 or Exact Match at top-1, is a strict evaluation metric that only assigns a score if the generated response perfectly matches the reference answer, without any deviations. Even if the response is semantically correct but differs in phrasing, formatting, or minor details, the model is penalized. In contrast, EM@1-Refined offers a more flexible evaluation, allowing for slight variations in wording.

The LLM-Match metric, introduced in [9], uses GPT-4 to assign a score from 1 to 5 to the generated responses based on their similarity to the reference answers. A score of 1 indicates a completely incorrect response, while a score of 5 corresponds to a fully correct answer. Intermediate scores reflect varying degrees of alignment.

4. More Visualization on ScanQA

In Fig. 1 and Fig. 2, we show more qualitative examples on ScanQA. Our SplatTalk can reason about the spatial relationships between objects in the scene significantly better than LLaVA-OV in many cases.

References

[1] 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. 1

[2] 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. 2

[3] An Yang et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 1

[4] 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. 2

[5] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. 1

[6] 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

[7] 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. 2

[8] 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

[9] 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. 2

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[11] 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. 1

[12] 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. 2

[13] 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. 1

[14] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1

[15] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging

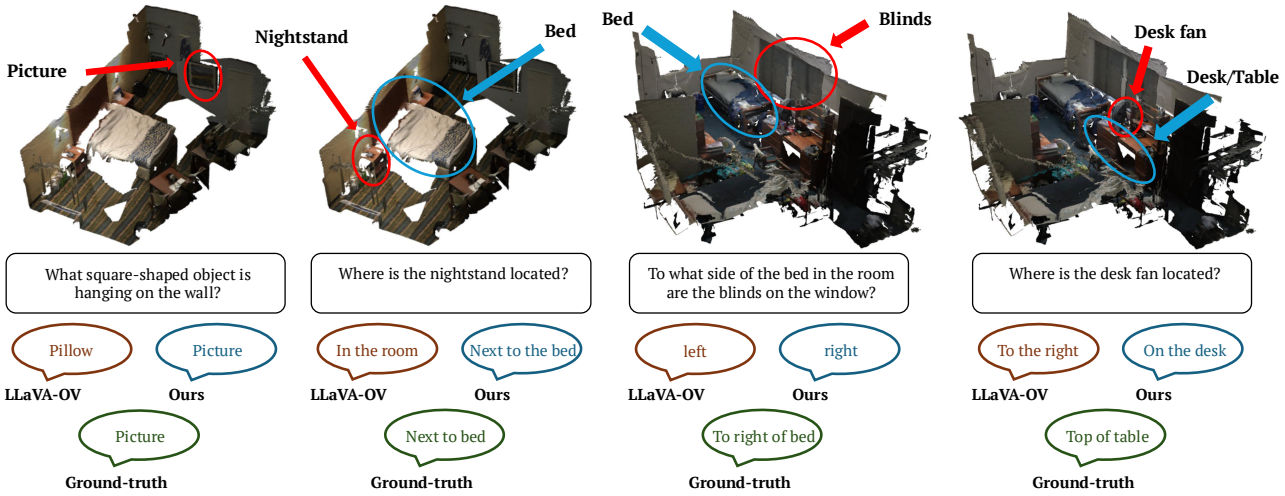

Figure 1. **Qualitative results on ScanQA**, scene0389_00 and scene0222_00. We compare responses from LLaVA-OV, our model (Ours), and the ground truth (GT) for spatial reasoning questions in 3D VQA. Each scene highlights the referenced objects with red circles, and key relative objects are marked in 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.

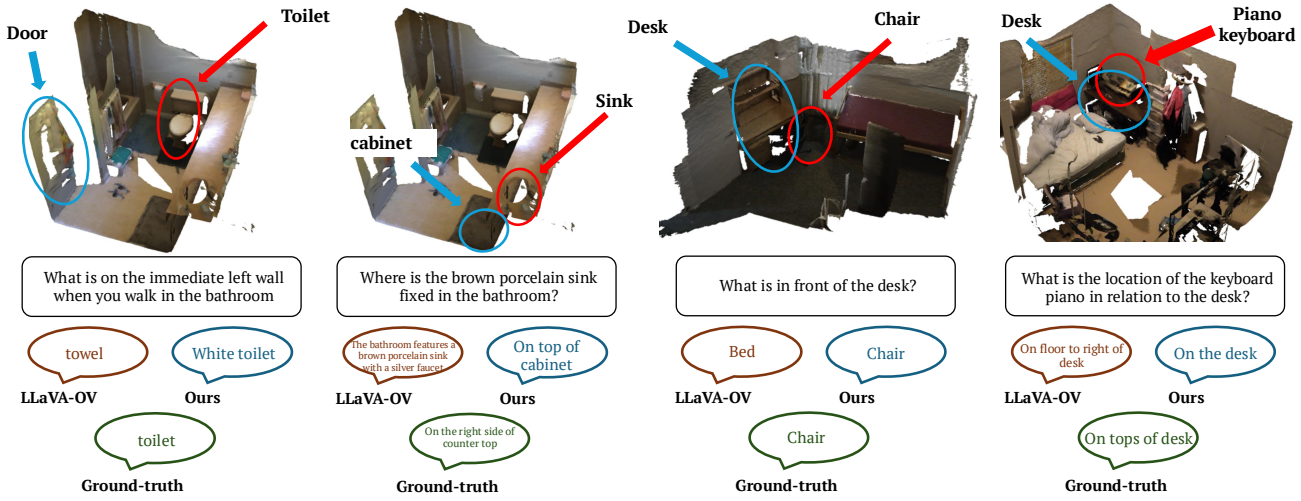

Figure 2. **Qualitative results on ScanQA**, scene0100_00, scene0193_00, and scene0426_00. We compare responses from LLaVA-OV, our model (Ours), and the ground truth (GT) for spatial reasoning questions in 3D VQA. Each scene highlights the referenced objects with red circles, and key relative objects 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.

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. 1