# Appendix

## A. Implementation Details

**Dataset Statistics.** We present the detailed statistics for training and testing data in Table 6 and 7, respectively. Following previous work [17, 18], we adopt the validation set for ScanRefer [5], Multi3DRefer [46], Scan2Cap [6], ScanQA [6], and the test set for SQA3D [31]. All data have been converted to LLaVA [29] format, and we conduct statistics in this format.

**Evaluation Details.** For ScanRefer [5], we select the object with the highest similarity as the prediction. For Multi3DRefer [46], we select objects with the highest probabilities such that their cumulative probability exceeds a given threshold p, which is empirically set to 0.25. For Scan2Cap [6], we follow [18] to evaluate the captioning performance by inserting "sos" and "eos" at the start and end of the prediction, respectively. Responses are generated using greedy sampling for 3D dense captioning and 3D question answering tasks.

## B. Detailed Comparison

**SQA3D.** We conduct a detailed evaluation on the test split of the SQA3D [31] dataset. As shown in Table 8, our model achieves the best performance on all categories of questions with an average EM at 58.86%, outperforming the previous state-of-the-art method LLaVA-3D [51] by 2.94% on the average EM.

**Scan2Cap.** As shown in Table 9, we provide a detailed comparison on the validation set of Scan2Cap [6]. During inference, we utilize the object proposals from [18], which include 50 predicted objects extracted with Mask3D [34] for each scan. From the table, we can see our method achieves state-of-the-art results on CIDEr and BLEU-4 at 83.77 and 42.43, respectively.

**ScanRefer.** We present a detailed comparison for ScanRefer [5] in Table 10. The table shows that our method reaches a peak of 58.12% Acc@0.25 and 51.72% Acc@0.5, surpassing ChatScene [17] by 2.6% and 1.5%, respectively.

**Multi3DRefer.** We follow previous work [46] to report the metrics across all question types, where "ZT" denotes zero-target, "ST" denotes single-target, "MT" denotes multi-target, "w/ D" and "w/o D" denote 'with and without distractors, respectively. As shown in Table 11, our method outperforms previous methods on "ZT w/o D", "ZT w/ D", and "ST w/D" types. However, the performance for "MT" is lower than ChatScene [17], suggesting that our method still struggles to distinguish similar objects.

**ScanQA.** We test our model on the validation set of ScanQA [2]. Compared to previous top-tier models, our Video-3D LLM achieves a relative improvement of 10.7% and 11.9% on EM@1 and CIDEr, respectively.

| | Data Count | Scan Count | Ques length | Answer Length |
|---|---|---|---|---|
| ScanRefer [5] | 36,665 | 562 | 24.9 | – |
| Multi3DRefer [46] | 43,838 | 562 | 34.8 | – |
| Scan2Cap [6] | 36,665 | 562 | 13.0 | 17.9 |
| ScanQA [2] | 26,515 | 562 | 13.7 | 2.4 |
| SQA3D [31] | 79,445 | 518 | 37.8 | 1.1 |

Table 6. Detailed statistics for training data. We report the average lengths for questions and answers, respectively.

| | Data Count | Scan Count | Ques length | Answer Length |
|---|---|---|---|---|
| ScanRefer [5] (Val) | 9,508 | 141 | 25.0 | – |
| Multi3DRefer [46] (Val) | 11,120 | 141 | 34.7 | – |
| Scan2Cap [6] (Val) | 2,068 | 141 | 13.0 | 18.7 |
| ScanQA [2] (Val) | 4,675 | 71 | 13.8 | 2.4 |
| SQA3D [31] (Test) | 3,519 | 67 | 36.3 | 1.1 |

Table 7. Detailed statistics for testing data. We report the average lengths for questions and answers, respectively.

| Method | Test set | | | | | | **Avg.** |
|---|---|---|---|---|---|---|---|
| | What | Is | How | Can | Which | Others | |
| SQA3D [31] | 31.6 | 63.8 | 46.0 | 69.5 | 43.9 | 45.3 | 46.6 |
| 3D-VisTA [52] | 34.8 | 63.3 | 45.4 | 69.8 | 47.2 | 48.1 | 48.5 |
| LLaVA-Video[47] | 42.7 | 56.3 | 47.5 | 55.3 | 50.1 | 47.2 | 48.5 |
| Scene-LLM [15] | 40.9 | 69.1 | 45.0 | 70.8 | 47.2 | 52.3 | 54.2 |
| LEO [18] | – | – | – | – | – | – | 50.0 |
| ChatScene [17] | 45.4 | 67.0 | 52.0 | 69.5 | 49.9 | 55.0 | 54.6 |
| LLaVA-3D [51] | – | – | – | – | – | – | 55.6 |
| Video-3D LLM (Uniform) | 51.1 | 72.4 | 55.5 | 69.8 | 51.3 | 56.0 | **58.6** |
| Video-3D LLM (MC) | 50.0 | 70.7 | 57.9 | 69.8 | 50.1 | 55.8 | 57.7 |

Table 8. Performance comparison on the test set of SQA3D [31].

| Method | @0.5 | | | |
|---|---|---|---|---|
| | C | B-4 | M | R |
| Scan2Cap [6] | 39.08 | 23.32 | 21.97 | 44.48 |
| 3DJCG [3] | 49.48 | 31.03 | 24.22 | 50.80 |
| D3Net [7] | 62.64 | 35.68 | 25.72 | 53.90 |
| 3D-VisTA [52] | 66.9 | 34.0 | 27.1 | 54.3 |
| LL3DA [9] | 65.19 | 36.79 | 25.97 | 55.06 |
| LEO [18] | 68.4 | 36.9 | 27.7 | 57.8 |
| ChatScene [17] | 77.19 | 36.34 | 28.01 | 58.12 |
| LLaVA-3D [51] | 79.21 | 41.12 | **30.21** | **63.41** |
| Video-3D LLM (Uniform) | **83.77** | **42.43** | 28.87 | 62.34 |
| Video-3D LLM (MC) | 80.00 | 40.18 | 28.49 | 61.68 |

Table 9. Performance comparison on the validation set of Scan2Cap [6]. C, B-4, M, R represent CIDEr, BLEU-4, Meteor, Rouge-L, respectively.

| Method | Venue | Unique | | Multiple | | Overall | |
|---|---|---|---|---|---|---|---|
| | | Acc@0.25 | Acc@0.5 | Acc@0.25 | Acc@0.5 | Acc@0.25 | Acc@0.5 |
| ScanRefer [5] | ECCV20 | 76.33 | 53.51 | 32.73 | 21.11 | 41.19 | 27.40 |
| MVT [19] | CVPR22 | 77.67 | 66.45 | 31.92 | 25.26 | 40.80 | 33.26 |
| 3DVG-Transformer [48] | ICCV21 | 81.93 | 60.64 | 39.30 | 28.42 | 47.57 | 34.67 |
| ViL3DRel [8] | NeurIPS22 | 81.58 | 68.62 | 40.30 | 30.71 | 47.94 | 37.73 |
| 3DJCG [3] | CVPR22 | 83.47 | 64.34 | 41.39 | 30.82 | 49.56 | 37.33 |
| D3Net [7] | ECCV22 | – | 72.04 | – | 30.05 | – | 37.87 |
| M3DRef-CLIP [46] | ICCV23 | 85.3 | 77.2 | 43.8 | 36.8 | 51.9 | 44.7 |
| 3D-VisTA [52] | ICCV23 | 81.6 | 75.1 | 43.7 | 39.1 | 50.6 | 45.8 |
| 3D-LLM (Flamingo) [16] | NeurIPS23 | – | – | – | – | 21.2 | – |
| 3D-LLM (BLIP2-flant5) [16] | NeurIPS23 | – | – | – | – | 30.3 | – |
| Grounded 3D-LLM [10] | ArXiv24 | – | – | – | – | 47.9 | 44.1 |
| PQ3D [53] | ECCV24 | 86.7 | 78.3 | 51.5 | 46.2 | 57.0 | 51.2 |
| ChatScene [17] | NeurIPS24 | 89.59 | 82.49 | 47.78 | 42.90 | 55.52 | 50.23 |
| LLaVA-3D [51] | ArXiv24 | – | – | – | – | 54.1 | 42.2 |
| Video-3D LLM (Uniform) | – | 87.97 | 78.32 | 50.93 | 45.32 | **58.12** | **51.72** |
| Video-3D LLM (MC) | – | 86.61 | 77.02 | 50.95 | 44.96 | 57.87 | 51.18 |

Table 10. Performance comparison on the validation set of ScanRefer [5]. "Unique" and "Multiple" depends on whether there are other objects of the same class as the target object.

| Method | ZT w/o D | ZT w/ D | ST w/o D | | ST w/ D | | MT | | ALL | |
|---|---|---|---|---|---|---|---|---|---|---|
| | F1 | F1 | F1@0.25 | F1@0.5 | F1@0.25 | F1@0.5 | F1@0.25 | F1@0.5 | F1@0.25 | F1@0.5 |
| M3DRef-CLIP [46] | 81.8 | 39.4 | 53.5 | 47.8 | 34.6 | 30.6 | 43.6 | 37.9 | 42.8 | 38.4 |
| D3Net [7] | 81.6 | 32.5 | – | 38.6 | – | 23.3 | – | 35.0 | – | 32.2 |
| 3DJCG [3] | 94.1 | 66.9 | – | 26.0 | – | 16.7 | – | 26.2 | – | 26.6 |
| Grounded 3D-LLM [10] | – | – | – | – | – | – | – | – | 45.2 | 40.6 |
| PQ3D [53] | 85.4 | 57.7 | – | 68.5 | – | 43.6 | – | 40.9 | – | 50.1 |
| ChatScene [17] | 90.3 | 62.6 | 82.9 | 75.9 | 49.1 | 44.5 | 45.7 | 41.1 | 57.1 | 52.4 |
| Video-3D LLM (Uniform) | 94.7 | 78.5 | 82.6 | 73.4 | 52.1 | 47.2 | 40.8 | 35.7 | **58.0** | **52.7** |
| Video-3D LLM (MC) | 94.1 | 76.7 | 81.2 | 72.6 | 52.7 | 47.4 | 40.6 | 35.3 | 57.9 | 52.4 |

Table 11. Performance comparison on the validation set of Multi3DRefer [46]. ZT: zero-target, ST: single-target, MT: multi-target, D: distractor.

| Method | Venue | EM | B-1 | B-2 | B-3 | B-4 | ROUGE-L | METEOR | CIDEr |
|---|---|---|---|---|---|---|---|---|---|
| ScanQA [2] | CVPR22 | 21.05 | 30.24 | 20.40 | 15.11 | 10.08 | 33.33 | 13.14 | 64.86 |
| 3D-VisTA [52] | ICCV23 | 22.4 | – | – | – | 10.4 | 35.7 | 13.9 | 69.6 |
| Oryx-34B [30] | ArXiv24 | – | 38.0 | 24.6 | – | – | 37.3 | 15.0 | 72.3 |
| LLaVA-Video-7B [47] | ArXiv24 | – | 39.71 | 26.57 | 9.33 | 3.09 | 44.62 | 17.72 | 88.70 |
| 3D-LLM (Flamingo) [16] | NeurIPS23 | 20.4 | 30.3 | 17.8 | 12.0 | 7.2 | 32.3 | 12.2 | 59.2 |
| 3D-LLM (BLIP2-flant5) [16] | NeurIPS23 | 20.5 | 39.3 | 25.2 | 18.4 | 12.0 | 35.7 | 14.5 | 69.4 |
| Chat-3D [40] | ArXiv23 | – | 29.1 | – | – | 6.4 | 28.5 | 11.9 | 53.2 |
| NaviLLM [49] | CVPR24 | 23.0 | – | – | – | 12.5 | 38.4 | 15.4 | 75.9 |
| LL3DA [9] | CVPR24 | – | – | – | – | 13.53 | 37.31 | 15.88 | 76.79 |
| Scene-LLM [15] | ArXiv24 | 27.2 | 43.6 | 26.8 | 19.1 | 12.0 | 40.0 | 16.6 | 80.0 |
| LEO [18] | ICML24 | – | – | – | – | 11.5 | 39.3 | 16.2 | 80.0 |
| Grounded 3D-LLM [10] | ArXiv24 | – | – | – | – | 13.4 | – | – | 72.7 |
| ChatScene [17] | NeurIPS24 | 21.62 | 43.20 | 29.06 | 20.57 | 14.31 | 41.56 | 18.00 | 87.70 |
| LLaVA-3D [51] | arXiv24 | 27.0 | – | – | – | 14.5 | **50.1** | **20.7** | 91.7 |
| Video-3D LLM (Uniform) | – | **30.10** | **47.05** | **31.70** | **22.83** | 16.17 | 49.02 | 19.84 | **102.06** |
| Video-3D LLM (MC) | – | 29.50 | 46.23 | 31.22 | 22.71 | **16.28** | 48.19 | 19.36 | 100.54 |

Table 12. Performance comparison on the validation set of ScanQA [2]. EM indicates exact match accuracy, and B-1, B-2, B-3, B-4 denote BLEU-1, -2, -3, -4, respectively.