3D-IDE: 3D Implicit Depth Emergent

Supplementary Material

6. Datasets Statistics

Training data. For fine-tuning, we adopt the same pool of 3D Understanding and Reasoning benchmarks as Video-3D LLM [51], namely ScanRefer, Multi3DRefer, Scan2Cap, ScanQA, and SQA3D. In total, this yields 223,128 training examples: SQA3D provides 79,445 samples (35.6% of the corpus), Multi3DRefer 43,838 (19.6%), ScanRefer and Scan2Cap 36,665 each (16.4% per dataset), and ScanQA 26,515 (11.9%). All datasets except SQA3D are built on 562 reconstructed scans, while SQA3D covers 518 scans. The average question length ranges from 13 to 38 words across datasets. Scan2Cap and ScanQA additionally offer answer sentences averaging 17.9 and 2.4 words, and SQA3D has relatively long questions (37.8 words on average) with very short answers (1.1 words).

Evaluation data. For evaluation, we use the official validation splits of ScanRefer, Multi3DRefer, Scan2Cap, and ScanQA, together with the test split of SQA3D. The combined evaluation suite contains 30,890 instances: 11,120 from Multi3DRefer (36.0%), 9,508 from ScanRefer (30.8%), 4,675 from ScanQA (15.1%), 3,519 from SQA3D (11.4%), and 2,068 from Scan2Cap (6.7%). The average question length in these splits varies between 13.0 and 36.3 words, while Scan2Cap and ScanQA provide answer texts with mean lengths of 18.7 and 2.4 words, respectively; SQA3D again features long questions (36.3 words) with very short answers (1.1 words).

7. Additional Ablative Analysis

Geometric Representation. As shown in Tabs. 11 and 12, both the pretrained-head and from-scratch-head variants improve encoder probing scores compared to training without any validator, confirming that geometric supervision is beneficial. Among them, the from-scratch validator yields the highest normal and correspondence accuracy, indicating that it encourages the encoder to internalize 3D structure more effectively than relying on a stronger pretrained head. This outcome is consistent with the design philosophy of IGEP: the geometric validator is intentionally kept weak and low-capacity so that it cannot absorb complex 3D reasoning on its own. To minimize the geometric loss, the visual encoder is instead pressured to internalize 3D structure within the shared tokens so that this low-capacity validator can decode it. Geometry therefore neither explicitly injected nor disentangled, but emerges under optimization pressure within a unified representation space. For surface normal estimation, we follow the Probe3D [17] protocol and train linear probing heads on features extracted from the frozen visual encoder. The reported geometry is thus derived entirely from the implicit RGB tokens; the training-only validator is not involved at evaluation time.

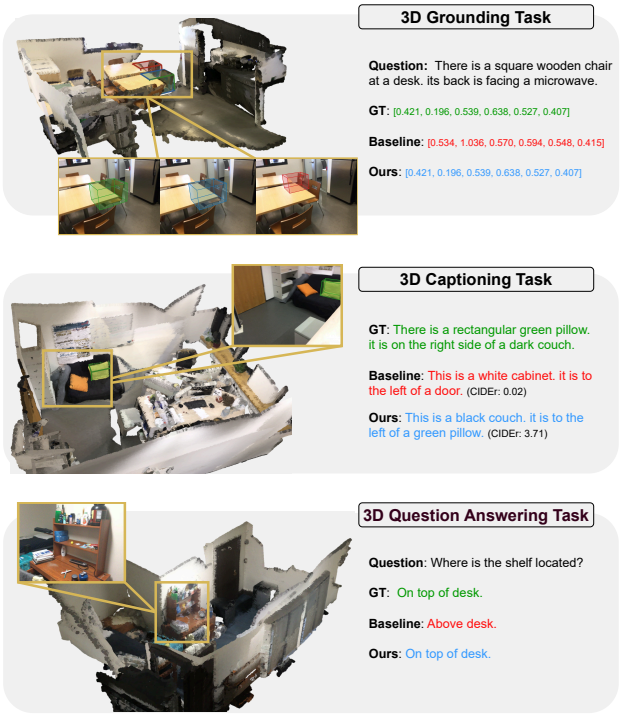

Figure 5. More qualitative results on three 3D vision-language tasks: language-guided object localization (top), region-level captioning (middle), and spatial question answering (bottom). In the grounding examples, green 3D bounding boxes denote the ground-truth targets, red boxes the predictions of the baseline, and blue boxes the predictions of our model. Our method better aligns with the targets and produces more accurate captions and answers.

Role of Global Supervision. A natural concern is that 3D-IDE might simply inherit 3D knowledge from the founda-

Table 6. **Additional Ablation Study.** Effect of $\mathcal{L}_{\text{global}}$ to speed-up training. ✓ denotes enabled, × denotes disabled.

Components			ScanRefer		Multi3DRef	
Global.	Geometric.	Cross-view.	Acc@0.25	Acc@0.5	F1@0.25	F1@0.5
×	×	×	53.7	47.8	46.0	42.4
✓	×	×	56.9	50.8	55.6	51.1
×	✓	✓	58.6	51.9	57.8	52.5
✓	✓ scratch	×	59.8	53.3	59.7	54.3
✓	✓ scratch	✓	**60.9**	**54.5**	**59.8**	**54.9**

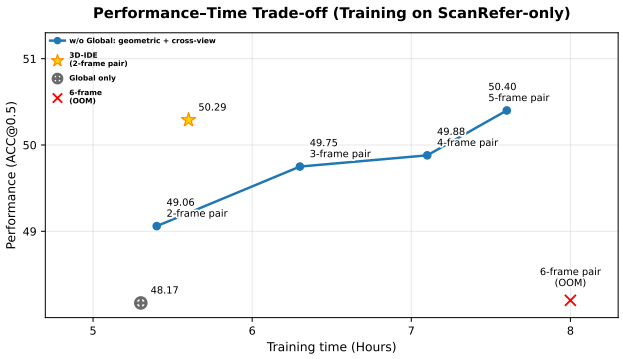

Figure 6. 3D-IDE retains performance w/o global supervision, at the cost of longer training and higher VRAM (OOM at 6-frame).

tion model f_G [38] through the global supervision. To disentangle this effect, we ablate $\mathcal{L}_{\text{global}}$ in Tab. 6. Using only local geometric and cross-view constraints, *i.e.*, $\mathcal{L}_{\text{geometry}}$ and $\mathcal{L}_{\text{cross-view}}$ with 2-frame warping, already yields a substantial gain over the baseline and even surpasses using $\mathcal{L}_{\text{global}}$ alone. This indicates that the principal source of 3D awareness arises from IGEP itself rather than being dictated by f_G. At the same time, combining local and global supervision achieves the best overall performance, while $\mathcal{L}_{\text{global}}$ is used at training and is entirely discarded at inference. In practice, the global term behaves as a training-time scene-level regularizer that approximates dense multi-view constraints whose pairwise complexity grows quadratically with sequence length, delivering an almost "free-lunch" improvement in 3D consistency without any inference latency. Importantly, f_G and the geometric validator operate on distinct parts of the architecture: f_G supervises only the final VLM hidden space via $\mathcal{L}_{\text{global}}$, whereas the geometric validator acts solely on the upstream shared encoder via $\mathcal{L}_{\text{geometry}}$ and $\mathcal{L}_{\text{cross-view}}$. This separation avoids direct coupling between the teacher and validator feature spaces, reducing the risk of misaligned supervision signals. As shown in Fig. 6, 3D-IDE retains strong performance even without global supervision, though at the cost of longer training and higher VRAM usage.

8. Detailed Comparison

Here, we conduct a thorough comparison with other methods, covering all metrics across five benchmark tasks.

Extended Ablation Across All Benchmarks. To further validate the contribution of each component, Tab. 7 extends the ablation in the main paper to all five benchmarks by cumulatively adding each objective. Each component brings consistent improvements across tasks, and the full configuration achieves the best results on all five benchmarks.

Impact of Removing 3D Inputs from 3DRS. A key claim of our work is that methods relying on ground-truth depth and camera pose at inference time suffer a severe performance drop when those inputs are withheld. Tab. 8 substantiates this by comparing 3DRS [24] in its original setting (with 3D geometric inputs) against its RGB-only variant (3DRS*, without 3D inputs) and our method, which is RGB-only by design. Removing 3D inputs causes a sharp drop in 3DRS performance on both ScanRefer and Multi3DRefer, whereas 3D-IDE remains strong and outperforms 3DRS* by a substantial margin on all grounding metrics. This confirms that 3DRS cannot be categorized as a generalist model without 3D geometric inputs, as it requires ground-truth depth and camera pose at inference to construct coordinate maps.

Components	ScanRefer Acc@0.25 ↑	Multi3DRefer F1@0.25 ↑	Scan2Cap CIDEr ↑	ScanQA CIDEr ↑	SQA3D EM ↑
Baseline	53.7	46.0	31.5	99.7	58.6
+ $\mathcal{L}_{\text{global}}$	56.9	55.6	77.7	100.0	57.8
+ $\mathcal{L}_{\text{geometry}}$	59.8	59.7	78.7	101.9	59.0
+ $\mathcal{L}_{\text{cross-view}}$ (Full)	**60.9**	**59.8**	**79.0**	**102.1**	**59.2**

Table 7. Extended ablation across all five benchmarks. Each row cumulatively adds one IGEP component. The full model consistently achieves the best results across all tasks.

Method	3D inputs	ScanRefer		Multi3DRefer	
		Acc@0.25	Acc@0.5	F1@0.25	F1@0.5
3DRS [24]	✓	62.9	56.1	60.4	54.9
3DRS* [24]	×	56.95	50.83	55.8	51.1
3D-IDE (Ours)	×	**60.9**	**54.5**	**59.8**	**54.9**

Table 8. Effect of removing 3D geometric inputs from 3DRS at inference. 3DRS* denotes the RGB-only variant. Our method closes most of the gap to the full 3DRS while using no 3D inputs.

ScanRefer. As shown in the detailed ScanRefer [5] results in Tab. 13, our method achieves strong overall performance, with clear improvements over the baseline on both Acc@0.25 and Acc@0.5, indicating better fine-grained localization of the target object.

Multi3DRefer. Following [47], we evaluate all question types, including zero-target (ZT), single-target (ST), and multi-target (MT) cases, with and without distractors. From Tab. 14, our approach consistently outperforms previous methods on the ST and MT splits under both distractor settings, demonstrating stronger robustness to spurious objects. Interestingly, the depth-free variants of Video 3D-LLM and 3DRS obtain higher ZT scores but substantially worse ST and MT results, indicating a bias toward predicting no target once geometric cues are removed.

ScanQA. On the ScanQA validation set [2], our method achieves better results than prior approaches on key metrics such as EM@1 and CIDEr, and is competitive on BLEU and METEOR, as shown in Tab. 15. These results highlight

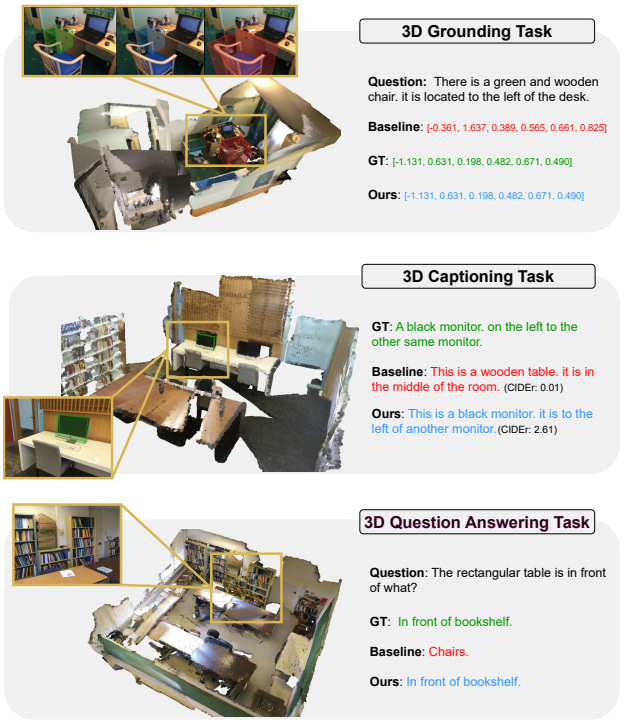

Figure 7. More qualitative results on three 3D vision-language tasks (continued). The three rows correspond to object localization, region-level captioning, and spatial question answering, respectively. Color coding is the same as in Figure 5. Our model remains effective across diverse scenes and linguistic queries.

the effectiveness of our model for 3D question answering.

SQA3D. As shown in Tab. 10, our method establishes new state-of-the-art performance on the SQA3D test split [32]. It attains the highest overall EM and consistently improves over previous methods across most question types, indicating strong generalization to diverse question categories.

Scan2Cap. On the Scan2Cap validation benchmark [13], we adopt the training and inference protocol of [54]. Under this setting, our method substantially improves over our baseline and attains CIDEr and BLEU-4 scores close to the best results, while remaining competitive on METEOR and ROUGE-L, as summarized in Tab. 9.

9. More Qualitative Results

Figs. 5 and 7 qualitatively summarize the behavior of our model on three challenging 3D scene understanding tasks: language-guided object localization, region-level captioning, and spatial question answering. In the visual grounding examples, the model must retrieve the correct object in a cluttered 3D environment given a natural-language description. For each case we visualize three bounding boxes: green denotes the ground-truth target, red the prediction of the RGB-only baseline, and blue the prediction of our model. Our predictions align much more closely with the intended targets, indicating that the model can reliably interpret both spatial and semantic cues from language. Across all three tasks, these qualitative results demonstrate that, even without any 3D input during inference, our method leverages its learned 3D-aware representation to produce more accurate and coherent outputs than the baseline.

Table 9. Performance comparison on the **Scan2Cap** validation set.

Method	@0.5			
	C	B-4	M	R
Scan2Cap [13]	39.08	23.32	21.97	44.48
3D-VisTA [57]	66.90	34.00	27.10	54.30
ChatScene [21]	77.19	36.34	28.01	58.12
LLaVA-3D [56]	79.21	41.12	30.21	63.41
baseline [51]	31.53	29.98	24.18	57.66
VG-LLM [53]	80.00	41.50	28.90	62.60
Ours	79.02	40.76	28.79	62.13

Method	Test set						**Avg.**
	What	Is	How	Can	Which	Others	
3D-VisTA [57]	34.8	63.3	45.4	69.8	47.2	48.1	48.5
Scene-LLM [18]	40.9	69.1	45.0	70.8	47.2	52.3	54.2
ChatScene [21]	45.4	67.0	52.0	69.5	49.9	55.0	54.6
LLaVA-3D [56]	–	–	–	–	–	–	55.6
Video-3D [51]	51.1	72.4	55.5	69.8	51.3	56.0	58.6
baseline [51]	51.8	73.1	56.5	70.1	51.0	54.7	58.5
Ours	**51.8**	72.7	**60.4**	68.3	49.0	**58.0**	**59.2**

Table 10. Performance comparison on the test set of **SQA3D**.

Table 11. **Correspondence Estimation Results for NAVI.** We present the NAVI correspondence estimation results for all models. The results are presented for features extracted at different layers with performance binned for different relative viewpoint changes between image pairs. The highest performing entry in each column is bolded.

		Block_0				Block_1				Block_2				Block_3			
Model	**Venue**	θ_0^{30}	θ_{30}^{60}	θ_{60}^{90}	θ_{90}^{120}	θ_0^{30}	θ_{30}^{60}	θ_{60}^{90}	θ_{90}^{120}	θ_0^{30}	θ_{30}^{60}	θ_{60}^{90}	θ_{90}^{120}	θ_0^{30}	θ_{30}^{60}	θ_{60}^{90}	θ_{90}^{120}
Video-3D LLM [54]	CVPR'25	**75.99**	**38.83**	**20.27**	10.68	80.48	52.97	32.15	17.70	75.36	49.26	34.94	21.38	71.97	46.28	34.50	22.19
3DRS [24]	NIPS'25	74.05	37.67	19.81	10.60	79.59	52.13	**33.07**	**17.87**	73.88	48.77	35.76	21.89	69.64	45.57	34.86	22.69
3D-IDE (Ours)$_{\text{pretrain}}$	–	74.63	38.07	19.99	10.69	81.60	53.64	33.03	17.33	77.04	**52.25**	37.13	22.34	72.33	47.68	35.15	22.11
3D-IDE (Ours)$_{\text{scratch}}$	–	74.93	38.05	19.89	**10.77**	**81.63**	**53.79**	33.06	17.38	**77.05**	52.02	**37.58**	**23.06**	**72.46**	**47.97**	**36.16**	**23.21**

Table 12. **Surface Normal Estimation Results.** We present the surface normal estimation results for all models. Higher is better for accuracy, lower is better for RMSE. The highest performing entry in each column is bolded.

		NAVI (Test)				
Model	**Venue**	Acc@11.25° (%)	Acc@22.5° (%)	Acc@30° (%)	RMSE (°) ↓	mAcc (%) ↑
Video-3D LLM [54]	CVPR'25	28.69	57.65	70.06	32.26	52.13
3DRS [24]	NIPS'25	28.61	57.90	70.48	31.86	52.33
3D-IDE (Ours)$_{\text{pretrain}}$	–	29.85	58.73	71.05	31.46	53.21
3D-IDE (Ours)$_{\text{scratch}}$	–	**30.15**	**59.01**	**71.32**	**31.46**	**53.49**

Table 13. **Performance comparison on the validation set of ScanRefer.** "Unique" and "Multiple" depends on whether there are other objects of the same class as the target object. 211

Method	Unique		Multiple		Overall	
	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
ScanRefer [5]	76.3	53.5	32.7	21.1	41.2	27.4
MVT [23]	77.7	66.4	31.9	25.3	40.8	33.3
3DVG-Transformer [50]	81.9	60.6	39.3	28.4	47.6	34.7
ViL3DRel [8]	81.6	68.6	40.3	30.7	47.9	37.7
3DJCG [3]	83.5	64.3	41.4	30.8	49.6	37.3
D3Net [6]	–	72.0	–	30.1	–	37.9
M3DRef-CLIP [47]	85.3	77.2	43.8	36.8	51.9	44.7
3D-VisTA [57]	81.6	75.1	43.7	39.1	50.6	45.8
3D-LLM (Flamingo) [20]	–	–	–	–	21.2	–
3D-LLM (BLIP2-flant5) [20]	–	–	–	–	30.3	–
Grounded 3D-LLM [11]	–	–	–	–	47.9	44.1
PQ3D [58]	86.7	78.3	51.5	46.2	57.0	51.2
ChatScene [21]	89.6	82.5	47.8	42.9	55.5	50.2
LLaVA-3D [56]	–	–	–	–	54.1	42.2
Video-3D LLM [51]	88.0	78.3	50.9	45.3	58.1	51.7
baseline [51]	82.17	73.71	45.10	40.14	52.29	46.66
3DRS* [24]	82.76	73.50	50.74	45.37	56.95	50.83
3D-IDE (Ours)	**86.72**	**77.94**	**54.73**	**48.90**	**60.94**	**54.53**

Table 14. **Performance comparison on Multi3DRefer validation set.** ZT: zero-target, ST: single-target, MT: multi-target, D: distractor.

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 [47]	81.8	39.4	53.5	47.8	34.6	30.6	43.6	37.9	42.8	38.4
D3Net [6]	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 [11]	–	–	–	–	–	–	–	–	45.2	40.6
PQ3D [58]	85.4	57.7	–	68.5	–	43.6	–	40.9	–	50.1
ChatScene [21]	90.3	62.6	82.9	75.9	49.1	44.5	45.7	41.1	57.1	52.4
Video-3D LLM [51]	94.7	78.5	82.6	73.4	52.1	47.2	40.8	35.7	58.0	52.7
3DRS [24]	95.6	79.4	79.6	71.4	57.0	51.3	43.0	37.8	60.4	54.9
baseline [51]	98.7	91.5	60.5	54.9	36.9	33.8	35.9	31.6	45.9	42.3
3DRS* [24]	96.6	85.2	75.1	67.4	49.0	44.8	42.6	37.6	55.8	51.1
3D-IDE (Ours)	95.6	79.7	**79.9**	**72.6**	**54.7**	**49.7**	**45.3**	**40.5**	**59.8**	**54.9**

Table 15. **Performance comparison on the validation set of ScanQA.** EM indicates exact match accuracy, and B-1, B-2, B-3, B-4 denote BLEU-1, -2, -3, -4, respectively.

Method	EM	B-1	B-2	B-3	B-4	ROUGE-L	METEOR	CIDEr
ScanQA [2]	21.05	30.24	20.40	15.11	10.08	33.33	13.14	64.86
3D-VisTA [57]	22.40	–	–	–	10.40	35.70	13.90	69.60
Oryx-34B [31]	–	38.00	24.60	–	–	37.30	15.00	72.30
LLaVA-Video-7B [49]	–	39.71	26.57	9.33	3.09	44.62	17.72	88.70
3D-LLM (Flamingo) [20]	20.40	30.30	17.80	12.00	7.20	32.30	12.20	59.20
3D-LLM (BLIP2-flant5) [20]	20.50	39.30	25.20	18.40	12.00	35.70	14.50	69.40
Chat-3D [41]	–	29.10	–	–	6.40	28.50	11.90	53.20
NaviLLM [52]	23.00	–	–	–	12.50	38.40	15.40	75.90
LL3DA [10]	–	–	–	–	13.53	37.31	15.88	76.79
Scene-LLM [18]	27.20	43.60	26.80	19.10	12.00	40.00	16.60	80.00
LEO [22]	–	–	–	–	11.50	39.30	16.20	80.00
Grounded 3D-LLM [11]	–	–	–	–	13.40	–	–	72.70
ChatScene [21]	21.62	43.20	29.06	20.57	14.31	41.56	18.00	87.70
LLaVA-3D [56]	27.00	–	–	–	14.50	50.10	20.70	91.70
Video 3D-LLM [51]	30.10	47.05	31.70	22.83	16.17	49.02	19.84	102.0
baseline [51]	29.5	46.9	31.3	22.7	16.2	48.8	19.6	100.5
3DRS* [24]	29.7	**47.9**	32.5	**23.8**	16.9	48.3	20.2	101.3
3D-IDE (Ours)	**29.8**	47.5	**32.9**	23.7	**17.4**	**48.8**	**20.8**	**102.1**