Geometry-Guided 3D Visual Token Pruning for Video-Language Models

Supplementary Material

A. Dataset Details

ScanRefer [3]. ScanRefer [3] is a large-scale 3D visual grounding dataset built on ScanNet [7], providing natural-language descriptions of indoor scenes to support the task of localizing target objects in 3D environments based on free-form textual queries. The dataset contains 51,583 descriptions from 800 ScanNet [7] scenes, covering 11,046 annotated objects, with an average of 13.81 objects and 64.48 descriptions per scene. The annotations are diverse and semantically rich, spanning more than 250 indoor object categories and frequently incorporating spatial language (98.7%), color (74.7%), and shape terms (64.9%). Following the official ScanNet [7] split, the dataset is divided into 36,665 training, 9,508 validation, and 5,410 test samples, with no scene overlap across splits. Evaluation is performed using thresholded Intersection over Union (IoU) accuracy, where a prediction is considered correct if its IoU with the ground-truth bounding box exceeds 0.25 or 0.5, offering a reliable measure of grounding quality.

Multi3DRefer [46]. Multi3DRefer [46] extends ScanRefer [3] to support grounding tasks involving a flexible number of target objects in real-world 3D scenes. The dataset includes 61,926 descriptions covering 11,609 objects, of which 51,583 are inherited from ScanRefer [3], 6,688 describe zero-target situations, and 13,178 involve multiple target objects. Scenes with multiple targets are typically offices or meeting rooms containing many chairs and tables. To evaluate grounding performance under variable target counts, Multi3DRefer [46] employs F1 scores at IoU thresholds of 0.25 and 0.5. During evaluation, per-pair IoUs between predicted and ground-truth boxes are computed, and the Hungarian algorithm is applied to obtain an optimal one-to-one matching. Pairs with IoUs exceeding the threshold are considered true positives. For zero-target cases, recall is set to 1, while precision is set to 1 if no predictions are made, and 0 otherwise.

Scan2Cap [6]. Scan2Cap [6] is a 3D dense captioning dataset built upon ScanRefer [3], targeting the joint task of object detection and natural-language description generation in 3D scenes. Each description provides information about the appearance of the object (*e.g.*, "a black wooden chair") as well as its spatial relationships with surrounding objects (*e.g.*, "the chair is placed at the end of the long dining table, just before the TV mounted on the wall"). Following the ScanRefer [3] split, the dataset contains 36,665 training samples and 9,508 validation samples. To jointly evaluate the quality of generated descriptions and the accuracy of detected bounding boxes, descriptions are assessed using standard image captioning metrics such as CIDEr [26] and BLEU-4 [19], combined with IoU scores between predicted and target bounding boxes.

ScanQA [1]. ScanQA [1] is a large-scale 3D question answering dataset built on ScanNet [7], designed to evaluate spatial understanding in 3D indoor scenes. In this task, models receive a full 3D reconstructed scene and answer free-form textual questions regarding objects and their spatial relations. The dataset contains 41,363 questions and 58,191 answers, including 32,337 unique questions and 16,999 unique answers. To account for variability in free-form responses, at least two ground-truth answers are provided for each validation and test question. Evaluation uses exact match (EM@K), which measures the proportion of samples for which any of the top-K predictions exactly matches one of the ground-truth answers. In addition, standard image captioning metrics, such as CIDEr [26], are applied to capture semantic similarity and enable robust evaluation for questions with multiple valid answer expressions.

SQA3D [18]. SQA3D [18] is a large-scale benchmark aimed at evaluating situated scene understanding in 3D environments. Given a 3D scene, the benchmark requires an agent to infer its situated context (*e.g.*, position and orientation) from a textual description, and subsequently reason about the surrounding environment to answer questions grounded in that situation. Built on 650 ScanNet [7] scenes, SQA3D [18] comprises 6.8k unique situations, 20.4k descriptive texts, and 33.4k diverse reasoning questions, covering spatial relations, navigation, common sense reasoning, and multi-hop reasoning. Following the ScanNet [7] data split, the dataset is divided into training, validation, and test sets. Evaluation relies on exact match (EM) accuracy as the primary metric, which is found to be sufficiently discriminative for comparing model performance.

B. More Quantitative Results

B.1. Comparison under the 8-Frame Video Setting

We additionally evaluate Geo3DPruner on a shorter video setting with 8-frame inputs, where each scene contains only 1,568 visual tokens. This reduced video context provides a weaker representation of the 3D scene, making token pruning more challenging. Tab. 6 reports the performance when retaining 640, 320, and 160 tokens, corresponding to pruning ratios of 60%, 80%, and 90%, respectively. Geo3DPruner preserves nearly all of the baseline performance at a pruning ratio of 60%, achieving 97.7% of the unpruned accuracy. Increasing the pruning ratio to 80% results in only a slight performance degradation, demon-

Table 6. **Performance comparison with previous visual token pruning methods when the video sequence length is set to 8.** *Avg.* represents the average percentage of performance maintained at the corresponding reduction ratio across five benchmarks and nine metrics. †: Following VG LLM [50], we introduce an additional 3D encoder to replace the original 3D positional embeddings.

Method	ScanRefer		Multi3DRefer		Scan2Cap		ScanQA		SQA3D	Avg.
	Acc@0.25	Acc@0.5	F1@0.25	F1@0.5	B-4@0.5	C@0.5	CIDEr	EM	EM	
8 Frames, All 1568 Tokens										
Video-3D LLM† [51]	52.1	46.3	51.5	47.0	39.9	76.0	99.2	28.0	58.0	100%
8 Frames, Retain 640 Tokens (↓60%)										
FastV [4]	48.7	43.1	48.1	43.9	35.4	62.7	93.0	26.7	56.8	92.4%
VisPruner [44]	49.7	44.3	49.2	45.0	35.5	62.3	94.1	27.5	56.9	93.8%
Geo3DPruner (Ours)	50.9	45.4	50.0	45.7	39.7	74.4	95.2	27.2	57.2	**97.7%**
8 Frames, Retain 320 Tokens (↓80%)										
FastV [4]	45.0	40.0	43.8	40.2	33.6	51.0	88.7	25.8	54.9	85.7%
VisPruner [44]	46.8	41.6	46.5	42.5	33.9	54.3	89.7	26.1	55.1	88.4%
Geo3DPruner (Ours)	48.4	43.1	47.1	43.3	39.1	70.6	92.5	26.5	56.7	**94.0%**
8 Frames, Retain 160 Tokens (↓90%)										
FastV [4]	41.5	36.7	37.7	34.6	32.4	45.1	83.4	23.9	53.5	78.7%
VisPruner [44]	42.8	38.0	40.6	37.2	33.5	48.6	84.9	24.3	53.4	81.6%
Geo3DPruner (Ours)	44.8	40.2	42.0	38.9	39.2	70.2	86.6	25.1	54.4	**88.7%**

Table 7. **Comparison with state-of-the-art methods** on 3D scene understanding benchmarks. The video sequence lengths for Video-3D LLM [51] and VG LLM [50] are set to 32. †: Following VG LLM [50], we introduce an additional 3D encoder to replace the original 3D positional embeddings.

Method	ScanRefer		Multi3DRefer		Scan2Cap		ScanQA		SQA3D
	Acc@0.25	Acc@0.5	F1@0.25	F1@0.5	B-4@0.5	C@0.5	CIDEr	EM	EM
ScanRefer [3]	37.3	24.3	-	-	-	-	-	-	-
M3DRef-CLIP [46]	51.9	44.7	42.8	-	38.4	-	-	-	-
Scan2Cap [6]	-	-	-	-	22.4	35.2	-	-	-
ScanQA [1]	-	-	-	-	-	-	64.9	21.1	47.2
3D-LLM (Flamingo) [10]	21.2	-	-	-	-	-	59.2	20.4	-
3D-LLM (BLIP2-FlanT5) [10]	30.3	-	-	-	-	-	69.4	20.5	-
LL3DA [5]	-	-	-	-	36.0	62.9	76.8	-	-
LEO [11]	-	-	-	-	38.2	72.4	101.4	21.5	50.0
LLaVA-3D [52]	54.1	42.4	-	-	41.1	79.2	91.7	27.0	55.6
Video-3D LLM [51]	58.1	51.7	58.0	52.7	41.3	83.8	102.1	**30.1**	58.6
VG LLM [50]	57.6	50.9	-	-	41.5	80.0	-	-	-
Video-3D LLM† [51]	**62.0**	**55.1**	**60.1**	**54.6**	**42.6**	**89.0**	**104.3**	29.8	**60.3**
+ Geo3DPruner (↓60%)	61.3	54.6	59.0	53.7	42.4	87.3	101.3	29.2	59.6
+ Geo3DPruner (↓80%)	60.1	53.5	57.5	52.5	42.1	85.7	97.9	28.3	58.3
+ Geo3DPruner (↓90%)	58.0	51.8	53.9	49.3	41.3	82.3	92.1	26.8	55.8

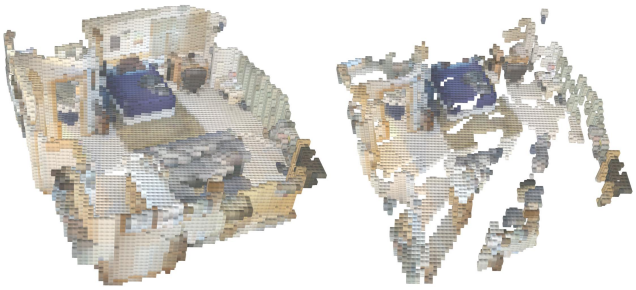
strating the robustness of our method even with limited visual information. With only 160 tokens retained under the extreme 90% pruning setting, Geo3DPruner maintains 88.7% of the baseline accuracy, substantially outperforming FastV [4] and VisPruner [44], which experience significant drops under the same conditions. These results indicate that even with limited 3D coverage, geometry-aware cross-frame correspondence enables effective token selection and robust scene understanding.

B.2. Comparison with State-of-the-Art Methods

We further evaluate the model pruned by Geo3DPruner against state-of-the-art methods, as summarized in Tab. 7. By incorporating geometry features, Video-3D LLM† [51] consistently outperforms existing methods across all benchmarks. After pruning 90% of the visual tokens, Video-3D LLM† [51] with Geo3DPruner maintains performance comparable to state-of-the-art methods while operating on a substantially reduced number of tokens.

Table 8. **Comparison under a fixed token budget** of 1568 tokens. The baseline uses 8 input frames without pruning, while Geo3DPruner leverages more frames with corresponding pruning ratios to maintain the same number of tokens. *Avg.* represents the average percentage of performance maintained at the corresponding reduction ratio across five benchmarks and nine metrics. †: Following VG LLM [50], we introduce an additional 3D encoder to replace the original 3D positional embeddings.

Method	ScanRefer		Multi3DRefer		Scan2Cap		ScanQA		SQA3D	Avg.
	Acc@0.25	Acc@0.5	F1@0.25	F1@0.5	B-4@0.5	C@0.5	CIDEr	EM	EM	
8 Frames, All 1568 Tokens										
Video-3D LLM† [51]	52.1	46.3	51.5	47.0	39.9	76.0	99.2	28.0	58.0	100%
16 Frames, Retain 1568 Tokens (↓50%)										
Geo3DPruner (Ours)	58.2	51.9	56.9	51.8	41.6	85.3	100.8	29.2	59.5	108%
32 Frames, Retain 1568 Tokens (↓75%)										
Geo3DPruner (Ours)	60.5	53.9	58.0	52.9	42.3	86.8	98.8	28.3	58.7	109%



(a) (b)

Figure 6. **Voxel-level visualization before and after pruning.** (a) Visualization of the original voxelized scene, where voxels densely cover both foreground objects and background regions. (b) Visualization after applying Geo3DPruner, where redundant voxels are largely removed while object-related and structurally important voxels are preserved.

B.3. Comparison under Fixed Token Budget

We further conduct a comparison under a fixed token budget of 1568 tokens. The baseline processes 8 frames without pruning, while our method utilizes more frames (16 and 32) combined with pruning ratios of 50% and 75%, respectively, to maintain the same number of visual tokens. As shown in Tab. 8, Geo3DPruner achieves substantially better performance on average, improving from 100% to 108% and 109% when using 16 and 32 frames, respectively. These results highlight a key advantage of our method, as it enables the model to process a larger number of frames while selectively retaining the most informative tokens. By leveraging geometry-guided pruning to remove cross-view redundancy, our method effectively increases the diversity of spatial observations under the same token budget, leading to more complete and robust 3D scene representations.

C. Voxel-level Visualization

To provide a more intuitive understanding of the proposed voxel-level pruning strategy, we visualize the 3D voxel representation of the scene before and after pruning. As shown in Fig. 6, the original voxel grid contains a large number of voxels distributed across both foreground objects and background regions. After applying Geo3DPruner, redundant voxels are effectively removed, while voxels corresponding to object-centric and structurally important regions are well preserved. This comparison clearly demonstrates that our method significantly reduces voxel redundancy while maintaining the overall spatial structure of the scene.