SEEING ONCE IS ENOUGH? ONLINE GEOMETRY-AWARE TOKEN PRUNING FOR 3D QUESTION ANSWERING

Ruei-Chi Lai[1] **Bolivar Solarte**[2] **Chin-Hsuan Wu**[3] **Yi-Hsuan Tsai**[4] **Min Sun**[1]
[1]National Tsing Hua University [2]Industrial Technology Research Institute ITRI
[3]University of Toronto [4]Atmanity Inc.

ABSTRACT

Recent Multi-modal Large Language Models (MLLMs) have demonstrated remarkable performance on 2D question answering tasks. However, extending these models to the 3D question answering remains challenging, as they typically require multiple views of the scene, which incurs substantial computational cost at inference. To mitigate this issue, existing solutions rely on strategic frame selection or token-merging algorithms that require preprocessing in advance all frames of the scene, i.e., an offline fashion. In contrast, we propose the first online token-pruning method that can be integrated seamlessly with current MLLM models for 3D question answering tasks, without additional training and with lower memory usage. Our key insight is to project each input frame into a shared voxel space using depth information and camera pose, identifying spatially-overlapped regions across frames and selectively pruning redundant image tokens before they enter the language model. Our method enables efficient online processing while reducing up to 50% of token usage. We apply this approach to Qwen2.5-VL-7B and Qwen3-VL-8B, demonstrating improved performance on the ScanQA, SQA3D, and OpenEQA-HM3D benchmarks.

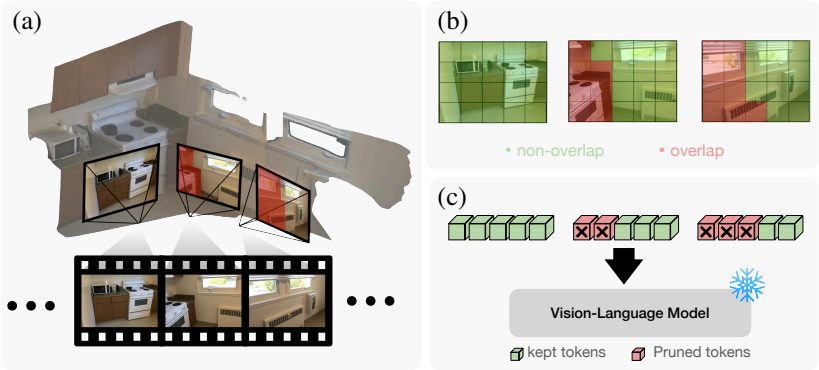

Figure 1: **Online token pruning.** (a) Sequentially received posed RGB-D frames are voxelized into a shared 3D space. (b) Overlapped regions across frames are identified through voxel comparison. (c) Redundant visual tokens corresponding to these regions are pruned prior to being fed into the VLM, enabling substantial reduction of token usage without compromising performance.

1 INTRODUCTION

Recent advances in Multi-modal Large Language Models (MLLMs) have demonstrated remarkable capabilities in 2D visual question answering (Bai et al., 2025; Team, 2025; Wang et al., 2025; Li et al., 2024; Achiam et al., 2023; Comanici et al., 2025). However, extending these models to 3D question answering remains challenging. Unlike 2D tasks that rely on a single image, 3D understanding requires processing multiple frames to obtain sufficient spatial information, which greatly increases computational costs and token usage.

Existing efforts have explored two main directions to address 3D question answering. First, several works inject 3D priors into MLLMs through specialized 3D modules to enhance spatial reasoning capabilities (Zheng et al., 2025; Xu et al., 2024b; Huang et al., 2025b; Chen et al., 2024a; Tang et al., 2024; Zhu et al., 2025; Hong et al., 2023; Zhang et al., 2024). However, these approaches face significant limitations, requiring specialized 3D modules, extensive 3D-language paired datasets, and costly retraining pipelines, which restrict their practical application. Another line of research aims to reduce visual token usage through strategic frame selection or token compression (Huang et al., 2025a; Zheng et al., 2025; Wu et al., 2025; Xu et al., 2024a). Although these approaches effectively reduce computational costs and maintain competitive performance, they require access to all images in the scene before inference, making them unsuitable for online tasks such as embodied AI, robotic navigation, or real-time scene understanding, which demand sequential, frame-by-frame processing. This raises the question: *How can we lower computational cost while preserving online processing?*

To this end, we propose the first online, training-free, geometry-aware token pruning method (Figure 1). Our approach leverages depth information and camera poses to project each input frame into a shared 3D voxel space, enabling the tracking of overlapping regions to identify and prune redundant visual tokens. Remarkably, we find that this geometry-aware pruning not only reduces computational costs but also improves performance on 3D question answering benchmarks, suggesting that removing redundant tokens helps the model to focus on more informative visual cues.

We validate our method on the ScanQA Azuma et al. (2022) SQA3D Ma et al. (2023) and OpenEQA-HM3D Arjun Majumdar (2024) benchmarks for 3D question answering tasks. We apply our online pruning method on two latest frontier Vision-Language Models (VLMs), i.e., Qwen2.5-VL-7B and Qwen3-VL-8B without fine-tuning their parameters. Across all experimental settings, our online pruning strategy reduces token usage by as much as **50%** and achieving an improvement of **+5.1** in the **LLM-Match** score. Our contributions are summarized as follows:

- We propose the first training-free, geometry-aware token pruning method for 3D question answering that can be seamlessly integrated into existing 2D VLMs and run in an online manner.
- Extensive experiments demonstrate that our pruning method can substantially reduce token usage while consistently improving performance across all settings.
- We further evaluate our approach across multiple models and benchmarks (ScanQA, SQA3D, OpenEQA-HM3D), demonstrating both its strong generalization capabilities and its effectiveness for 3D question answering tasks.

2 Related Work

2.1 3D MLLMs

With the rapid development of Multi-modal Large Language Models (MLLMs), numerous works (Chen et al., 2024a;b; Wang et al., 2023; Huang et al., 2024; Zhang et al., 2024; Qi et al., 2024; Zhu et al., 2025; Xu et al., 2024b; Zheng et al., 2025; Huang et al., 2025b) recently boost the MLLMs with 3D scene understanding capabilities using 3D information such as 3D point clouds and 3D bounding boxes. LL3DA (Chen et al., 2024a) and Grounded 3D-LLM (Chen et al., 2024b) leverage 3D detectors or segmentation modules to extract object-centric features for language reasoning. Chat3D (Wang et al., 2023), LEO (Huang et al., 2024), and Chat-Scene (Zhang et al., 2024) adopt a similar approach by encoding segmented 3D objects and fusing their features into large language models. GPT4Scene (Qi et al., 2024) constructs a bird's-eye-view (BEV) image by reconstructing the 3D scene for question answering. LLaVA-3D (Zhu et al., 2025) integrates 3D geometric position information into image patch embeddings with instruction tuning to align 2D and 3D modalities, while Video-3D LLM (Zheng et al., 2025) employs a 3D positional encoding module with post-instruction tuning for 3D scene understanding. However, these methods depend on 3D training data, which is scarcer than large-scale 2D image datasets. The scarcity of high-quality 3D resources poses a bottleneck for scaling 3D MLLMs, limiting their coverage of real-world scene variations and capacity to learn generalizable spatial reasoning, highlighting the need for more data-efficient approaches.

2.2 3D QUESTION ANSWERING WITH VLMS

Recent advancements in Vision-Language Models (VLMs) have extended multimodal reasoning from 2D images to multi-frame and 3D scene understanding (Liu et al., 2025; Li et al., 2024; Bai et al., 2025; Team, 2025; Comanici et al., 2025; Achiam et al., 2023; Lin et al., 2024). Benchmarks such as ScanQA (Azuma et al., 2022), SQA3D (Ma et al., 2023), and OpenEQA (Arjun Majumdar, 2024) evaluate 3D question answering, while ScanRefer (Chen et al., 2020) and Multi3DRefer (Zhang et al., 2023) focus on 3D grounding. Unlike grounding, 3D question answering requires spatial and semantic reasoning across multiple views to answer scene-level questions.

Several VLM families, including LLaVA (Li et al., 2024), Video-LLaVA (Lin et al., 2024), and Qwen-VL (Bai et al., 2025; Team, 2025), have demonstrated strong cross-modal understanding through large-scale instruction tuning. However, most existing VLMs remain inherently 2D-based, relying solely on image sequences without explicitly incorporating 3D geometric information, limiting their spatial reasoning in complex 3D environments. In this work, we bridge this gap by exploring a balance between utilizing 3D spatial information and leveraging the strong generalization capability of 2D VLMs for effective 3D scene question answering.

2.3 FRAME SAMPLING STRATEGY

To enable MLLMs to understand 3D scenes through powerful visual reasoning, visual-based 3D scene understanding typically relies on video frames as input. However, due to limited GPU memory, models can only process a subset of frames, even though a single 3D scene dataset may contain thousands of frames. A widely adopted approach is uniform frame sampling (Bai et al., 2025; Yang et al., 2025; Zheng et al., 2025; Huang et al., 2025a), which evenly selects frames from the entire sequence. While simple and efficient, this strategy often reaches its performance ceiling.

To overcome this limitation, several works (Zheng et al., 2025; Wu et al., 2025) have leveraged 3D geometric information and adopted greedy algorithms to select frames that maximize spatial coverage while minimizing the total number of frames. Some prior works (Hu et al., 2025) propose trainable frame selectors that predict frame importance scores for query-relevant frame identification. Similarly, VLM-Grounder (Xu et al., 2024a) introduces a query-aware frame selection framework, while Dynamic Token Compression (DTC) (Huang et al., 2025a) presents an offline token reduction method combining 3D voxelization with visual similarity matching to reduce visual tokens while maintaining performance.

However, most existing strategies rely on offline processing or additional selector modules, increasing computational overhead and making them unsuitable for online or real-time scenarios. We propose an online, single-pass frame sampling strategy that leverages 3D geometric information to reduce redundant tokens while improving model performance, effectively balancing efficiency and 3D scene understanding in 3D question answering tasks.

3 PROPOSED METHOD

In this section, we introduce our online geometry-aware token pruning method for the 3D question answering task, which aims to maximize 3D scene information while remaining computationally efficient. For reference purposes, the overall pipeline is illustrated in Figure 2. In Section 3.1, we discuss frame sampling strategies under both online and offline settings. In Section 3.2, we present our method to identify redundant information in overlapping regions under a sequential frame input setting. Finally, in Section 3.3, we describe how redundant information is dropped during MLLMs' inference without requiring any further post-training.

3.1 FRAME SAMPLING

Processing scene videos in VLMs requires selecting a subset of frames due to limited GPU memory and computational resources. An effective frame sampling strategy should preserve the essential spatial coverage of the 3D scene while minimizing redundancy among frames. Existing approaches typically perform offline frame selection, where the entire video or dataset must be available beforehand to optimize the coverage. However, this assumption is impractical for streaming or real-

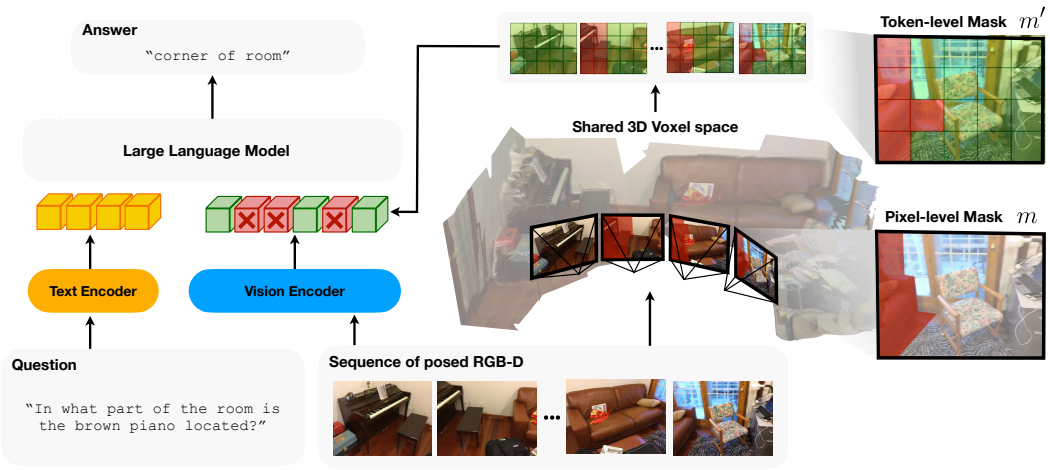

Figure 2: **Pipeline of our method.** Our method processes frames sequentially via online 3D voxelization, generating a pixel-level mask m from overlapped regions and computing a token-level mask m' to effectively prune redundant tokens before they enter the LLM.

time scenarios, where frames are observed sequentially and future content is unknown. To address this limitation, we first consider uniform sampling as a online strategy that does not require preprocessing the entire set of video frames. However, uniform sampling fails to account for spatial redundancy between frames and often leads to suboptimal scene coverage, motivating the need for a more adaptive online selection mechanism.

Uniform sampling. Uniform sampling intuitively selects frames from the video frame set $F = \{f_1, f_2, ..., f_n\}$ at a fixed interval of every n/t where t denotes the total number of frames to be sampled. However, uniform sampling introduces a critical trade-off between temporal coverage and computational efficiency. When the sampling interval is large, the model may skip important frames containing fine-grained spatial or semantic details of the 3D scene. In contrast, reducing the interval to capture more information leads to a larger number of selected frames, quickly exceeding GPU memory limits. As a result, uniform sampling often fails to achieve an optimal balance between efficiency and scene coverage.

Maximum coverage sampling. We follow the offline maximum coverage sampling strategy introduced in prior work (Zheng et al., 2025), which aims to select the fewest possible frames while maximizing the coverage of the 3D scene. This problem is formulated as a maximum coverage problem and is known to be NP-hard. Therefore, a greedy algorithm is adopted to obtain an approximate solution. Specifically, given a video with frame set $F = \{f_1, f_2, ..., f_n\}$, each frame is projected into 3D voxels, and the goal is to find a subset of frames $S \subseteq F$ that maximizes the overall voxel coverage.

3.2 3D Projection

For each incoming posed RGB-D image, we first project the image into the 3D world coordinate system, obtaining $C \in \mathbb{R}^{H \times W \times 3}$ for every pixel using the corresponding depth map and camera pose. Each pixel (u, v) is thereby associated with a 3D point $C(u, v) = (x, y, z)$ in the global coordinate space. The resulting 3D points are then voxelized to produce discrete voxel indices $\mathcal{V} \in \mathbb{Z}^{(H \cdot W) \times 3}$, which represent quantized spatial locations in the scene.

To determine whether a spatial region has been previously observed, we maintain a global voxel set $\mathbf{S}$ that records all voxels visited by past frames. For each new frame, voxels corresponding to $\mathcal{V}$ are compared against $\mathbf{S}$; if a voxel already exists in $\mathbf{S}$, it is considered overlapped. These overlapped voxels are subsequently back-projected onto the image plane to produce a pixel-level binary mask $m \in \{0, 1\}^{H \times W}$, where $m(u, v) = 1$ indicates that the corresponding pixel is projected from a region that has already been observed. This mask effectively captures spatial redundancy across frames at the pixel level.

To bridge the pixel-level redundancy and the visual tokens processed by the vision encoder, the mask m is spatially aggregated over non-overlapping patches of size $P \times P$ to form a token-level mask $m' \in \{0,1\}^{H' \times W'}$, where $H' = \lfloor \frac{H}{P} \rfloor$ and $W' = \lfloor \frac{W}{P} \rfloor$. Each patch corresponds to one visual token in the encoder, and we compute an overlap ratio for each token as:

$$r_{i,j} = \frac{1}{|\mathcal{B}i,j|} \sum_{(u,v) \in \mathcal{B}_{i,j}} m(u,v), \tag{1}$$

where $\mathcal{B}i,j$ denotes the set of pixels within the $(i,j)^{\text{th}}$ patch. The overlap ratio $r_{i,j}$ measures the proportion of pixels within a token region that have been previously observed, serving as a quantitative indicator of visual redundancy.

We then define the token-level binary mask as:

$$m'_{i,j} = \begin{cases} 1, & \text{if } r_{i,j} \geq \tau_o, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where τ_o is the pruning threshold that specifies the minimum proportion of overlap required for a token to be considered redundant. In practice, this threshold provides a controllable trade-off between pruning aggressiveness and information retention. A higher τ_o prunes only highly redundant tokens, while a lower value removes tokens more aggressively.

The resulting mask m' is later used to guide the pruning process of visual tokens before feeding them into the large language model. This mechanism ensures that only novel or spatially informative regions are retained, thereby reducing token redundancy and improving computational efficiency. Further analysis of the impact of different pruning thresholds is presented in the experimental section.

3.3 Token Pruning

The RGB image is first fed into the vision encoder to generate a grid of visual tokens $e \in \mathbb{R}^{H' \times W' \times d}$, where d denotes the token embedding dimension. Each token $e_{i,j}$ corresponds to a localized spatial region in the input frame and encodes its visual information.

To identify redundant information across frames, we leverage the 3D projection-based overlap analysis introduced in Section 3.2. The mask m' reflects which spatial regions have already been observed or overlapped by previously processed frames. Specifically, $m'_{i,j} = 1$ indicates that the corresponding token $e_{i,j}$ belongs to an overlapped region that conveys redundant visual content and should therefore be pruned. Conversely, $m'_{i,j} = 0$ denotes that the token originates from a novel or unobserved region and should be retained for feeding into the LLM. The pruned token set can thus be formulated as:

$$e' = \{\, e_{i,j} \mid m'_{i,j} = 0 \,\}, \tag{3}$$

where e' represents the subset of tokens that contribute unique spatial information. When feeding e' into the LLM, redundant tokens are excluded from both the prefill and forward phases, reducing computational overhead and mitigating attention dilution. This geometry-aware pruning removes visual tokens from already-explored regions, effectively reducing redundancy while preserving essential context. As a result, the VLM focuses more on novel spatial cues, improving both efficiency and 3D question answering performance.

4 Experiments

4.1 Experimental Setup

Implementation details. We use Qwen2.5-VL-7B and Qwen3-VL-8B as base models. Two sampling strategies are adopted: uniform sampling and maximum coverage (MC) sampling. For uniform sampling, we select 20 frames on the ScanQA (Azuma et al., 2022), SQA3D (Ma et al., 2023), and OpenEQA-HM3D (Arjun Majumdar, 2024) benchmarks. As discussed in Section 3.1, MC sampling adaptively selects frames, averaging around 20 per scene. To ensure a fair comparison in token usage, we uniformly sample 30 frames for all benchmarks before applying our method, while using the same selected frames for offline MC. The pruning threshold is set to 100%, images are resized to 640×480, and voxel size is 0.1 m. We will release the code to the public.

Datasets & evaluation metrics. We conduct our experiments on three 3D question answering datasets: ScanQA-val, SQA3D-test, and OpenEQA-HM3D to evaluate the effectiveness of our geometry-aware token pruning method. ScanQA and SQA3D are built upon the ScanNet dataset (Dai et al., 2017). Both emphasize evaluating MLLMs' spatial understanding capabilities in 3D scenes. ScanQA-val includes 71 distinct scenes and 4,675 questions, while SQA3D-test includes 67 distinct scenes and 3,519 questions. Different from the ScanNet dataset, OpenEQA-HM3D is built on the HM3D dataset, which comprises real-world questions and focuses on embodied question answering. OpenEQA-HM3D includes 63 distinct scenes and 557 questions.

In this study, we follow prior works (Azuma et al., 2022; Hong et al., 2023; Zheng et al., 2025; Wu et al., 2025) and report Exact Match and CIDEr scores for the ScanQA and SQA3D benchmarks. We also follow OpenEQA and report the GPT-4.1-mini LLM-Match score in our experiments. We further include the total token consumption required to evaluate the full benchmark, reported as "Tokens" in the table.

4.2 Experimental Results

Table 1: **Overall performance comparisons for the online strategy.** "Fine-tuned" indicates that the model is trained on the ScanQA and SQA3D datasets.

Model	Online Sampling	Sampling Strategy	ScanQA			SQA3D		OpenEQA-HM3D	
			EM ↑	CIDEr ↑	Tokens ↓	EM ↑	Tokens ↓	LLM Match ↑	Tokens ↓
Fine-tuned 3D LLMs									
Video-3D LLM	✓	Uniform	29.6	99.6	-	58.3	-	-	-
Zero-shot 2D VLMs									
Qwen2.5-VL-7B	✓	Uniform	24.1	65.6	37.1M	46.5	28.0M	53.1	**4.4M**
	✓	Uniform + Ours	**25.1**	**69.3**	**32.6M**	**47.3**	**24.1M**	**58.2**	5.2M
Qwen3-VL-8B	✓	Uniform	27.2	78.7	28.6M	50.2	21.5M	67.1	**3.4M**
	✓	Uniform + Ours	**27.9**	**79.9**	**26.2M**	**50.7**	**19.4M**	**67.8**	4.1M

Table 2: **Overall performance comparisons for offline strategy.** We report DTC EM score when retaining 54% of the tokens with 12 uniformly sampled frames. "Fine-tuned" indicates that the model is trained on the ScanQA and SQA3D datasets.

Model	Online Sampling	Sampling Strategy	ScanQA			SQA3D	
			EM ↑	CIDEr ↑	Tokens ↓	EM ↑	Tokens ↓
Fine-tuned 3D LLMs							
Video-3D LLM		MC	29.8	100.3	-	-	-
Zero-shot 2D VLMs							
LLava-OV-7B		DTC	27.8	-	-	-	-
Qwen2.5-VL-7B		MC	**26.2**	70.9	38.6M	**48.2**	28.5M
		MC + Ours	26.1	**71.4**	**28.3M**	47.8	**20.8M**
Qwen3-VL-8B		MC	28.0	80.1	29.8M	50.5	22.0M
		MC + Ours	**28.2**	**80.5**	**22.5M**	**51.1**	**16.6M**

We present the overall comparison between leveraging uniform sampling and maximum coverage sampling strategies, as well as applying our method, in both Table 2 and Table 1. Different from fine-tuned models that post-trained on downstream tasks, we use zero-shot VLMs such as Qwen2.5-VL-7B and Qwen3-VL-8B as our base models.

Online scenario. In Table 1, we compare our method with the online uniform sampling strategy, which uses 20 frames for ScanQA, SQA3D, and OpenEQA-HM3D. To ensure a fair comparison in token usage, we apply our method on 30 uniformly sampled frames for all benchmarks, while pruning overlapped visual tokens to match the baseline's token budget. As shown in the table, when

Table 3: **Categorical results on SQA3D.** Comparisons between online and offline sampling strategies, with evaluation results reported across six categories of SQA3D. Our method can be integrated with both strategies and enhances them.

Model	Online Sampling	Sampling Strategy	SQA3D what	which	can	is	how	others	Average ↑	Tokens ↓
Qwen2.5-VL-7B	✓	Uniform	41.2	45.9	**56.5**	**58.3**	37.4	45.8	46.5	28.0M
	✓	Uniform + Ours	**41.7**	**49.3**	53.6	57.8	**41.1**	**46.8**	**47.3**	**24.1M**
		MC	**42.0**	**49.6**	**56.2**	**58.1**	**43.0**	**48.1**	**48.2**	28.5M
		MC + Ours	41.8	49.0	55.3	**58.1**	41.7	47.9	47.8	**20.8M**
Qwen3-VL-8B	✓	Uniform	45.6	45.6	54.1	60.6	46.2	51.1	50.2	21.6M
	✓	Uniform + Ours	**46.4**	**48.2**	**55.6**	**60.9**	**44.3**	**51.4**	**50.7**	**19.4M**
		MC	46.7	41.9	**57.7**	61.2	46.0	50.5	50.5	22.0M
		MC + Ours	**47.2**	**42.5**	56.5	**61.8**	**47.1**	**51.9**	**51.1**	**16.6M**

applying our method with more frames, the redundant overlapped tokens identified by our pruning process are effectively removed. This allows the model to focus more on the informative regions, not only reducing token usage but also improving overall performance. The effectiveness of our method is consistently demonstrated on both QwenVL-series models, where performance improvements are observed across three benchmarks. Specifically, on Qwen2.5-VL-7B, our approach improves the Exact Match score from 24.1 to 25.1 on ScanQA and from 46.5 to 47.3 on SQA3D, while using fewer tokens. The results indicate that when the online setting is required, our method provides a more effective sampling strategy without introducing any preprocessing or selector modules that would increase computational overhead. To further analyze scalability, we present a detailed comparison of different numbers of uniformly sampled frames and the effect of applying our method in Section 4.3, as illustrated in Figure 3.

Offline scenario. In Table 2, we compare our method with the offline sampling strategies. We include prior work DTC (Huang et al., 2025a) in our comparison. DTC compresses visual tokens to reduce the input length, we report its EM score when retaining 54% of the tokens. In our evaluation setup, we adopt the adaptive maximum coverage (MC) sampling strategy, which selects approximately 20 frames per question across all benchmarks. As shown in the table, even under the same set of offline-processed frames—where MC is used to select the most representative and diverse frames—our method is able to significantly reduce token usage while maintaining competitive performance. This makes our approach particularly suitable for scenarios with limited computational resources. Furthermore, applying our method achieves comparable Exact Match scores on both ScanQA and SQA3D, improves the CIDEr score on ScanQA with Qwen2.5-VL-7B. These results demonstrate that our pruning strategy effectively removes redundant visual tokens, optimizing the input representation in a way that serves as an efficient offline enhancement to existing sampling methods.

Categorical results. In Table 3, we present the evaluation results across six categories of the SQA3D benchmark for both QwenVL-series models. For the online sampling strategies, the results demonstrate that our method not only improves the overall performance but also yields consistent gains across almost all categories. In the offline setting, our method maintains comparable performance across all six categories. Notably, for Qwen3-VL-8B, applying our method on top of the MC baseline achieves a 1.4 points improvement in Exact Match (EM) for the "others" category. Across both settings, our method consistently reduces the total number of visual tokens, highlighting its efficiency in balancing performance and token usage.

4.3 Ablation Study

Scaling up sampled frames. In Figure 3, We investigate the scalability of our method when increasing the number of sampled frames. From the figure, we observe that under the same 20-frame setting, our approach achieves better performance while using fewer visual tokens compared to the baseline uniform sampling strategy. As the number of frames increases to 30 and 50, our method consistently removes redundant tokens while maintaining or slightly improving Exact Match (EM)

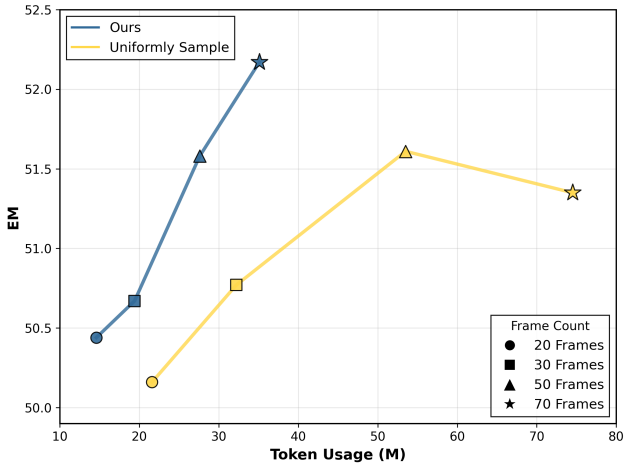

Figure 3: **Ablation on scaling the number of frames.** Our method maintains stable or improved performance while significantly reducing token usage compared to uniform sampling.

scores. When input frames scale up to 70, the baseline performance degrades due to excessive redundancy, while our method continues to yield stable or improved performance with significantly reduced token count. This indicates that our geometry-aware pruning effectively filters out repetitive visual information, allowing the model to focus on novel and spatially informative regions even when frame input becomes dense. This finding is particularly important in large-scale 3D environments, where comprehensive scene coverage often demands more input frames. While increasing frames typically causes quadratic growth in token usage and computational cost, our approach maintains a balance between accuracy and efficiency by pruning tokens while preserving critical scene context.

Table 4: **Ablation study on the pruning threshold τ_o.** We investigate how different threshold settings affect pruning behavior, with higher τ_o values leading to less aggressive token removal. The results demonstrate that a threshold of 100% performs the best.

Model	Sampling Strategy	τ_o	SQA3D	
			EM ↑	Tokens ↓
Qwen3-VL-8B	Uniform	-	51.4	74.5M
	Uniform + Ours	25%	49.6	**10.7M**
	Uniform + Ours	50%	50.8	13.5M
	Uniform + Ours	80%	50.8	19.4M
	Uniform + Ours	100%	**52.2**	35.2M

Pruning threshold. In Equation (1) and Equation (2), we introduce the pruning threshold τ_o, specifying the minimum proportion of overlapped pixels within a patch required for the corresponding image token to be considered redundant and pruned. This ensures that a token is removed only when a sufficiently large portion of its spatial region has already been observed. We evaluate the effect of different pruning thresholds on the SQA3D dataset, as shown in Table 4. Using 70 uniformly sampled frames as the baseline, we observe that the pruning threshold τ_o plays a crucial role in shaping the model's behavior. When a lower threshold is applied, model performance tends to decrease as more regions are considered redundant and aggressive pruning takes effect. In contrast, when the threshold reaches 100%—meaning all pixels within a patch are identified as overlapped—the model achieves better performance than the baseline, improving from 51.4 to 52.2 while using only 47% of the tokens. These results suggest that properly controlling the pruning threshold enables the system to effectively filter out redundant tokens, allowing the model to focus on novel and spatially informative visual regions rather than repeated observations.

4.4 QUALITATIVE RESULT

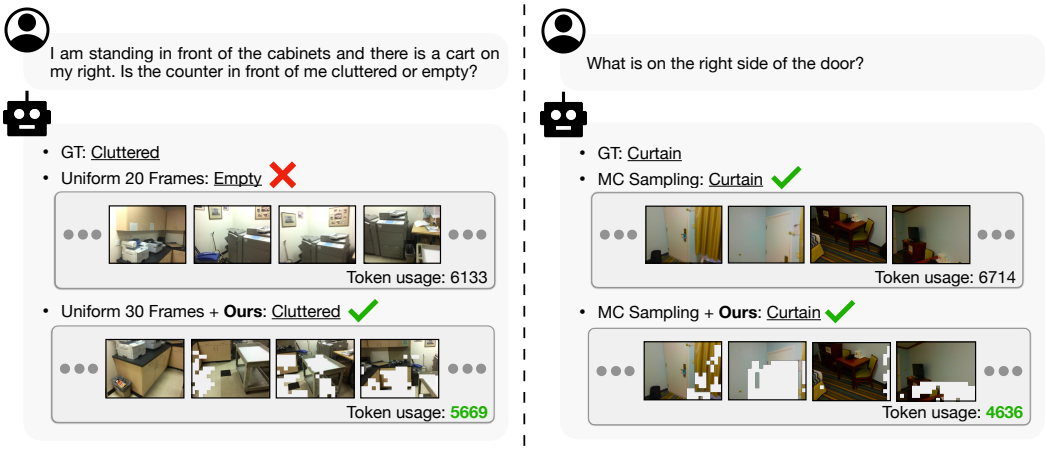

Figure 4: **Qualitative results.** We adopt our method with two sampling strategies: uniform (left) and MC (right), on SQA3D and ScanQA, respectively. Our method consistently improves token efficiency and performance.

In Figure 4, we present two visualization examples demonstrating our method's effectiveness under online uniform sampling and offline maximum coverage sampling. The left half compares uniform sampling with 20 frames against our method applied to 30 uniformly sampled frames. Despite processing more frames initially, our method significantly reduces the final token count through geometry-aware pruning. More importantly, this token reduction improves answer accuracy, while the baseline produces an incorrect response, our method successfully removes redundant visual information, allowing the model to focus on relevant spatial features and generate the correct answer. The right half shows our method combined with maximum coverage sampling. Both approaches produce correct answers, validating that our pruning preserves essential visual information. However, our method reduces token consumption from 6.7k to 4.6k tokens while maintaining identical performance. This substantial reduction in computational cost without sacrificing accuracy demonstrates that our method successfully identifies and eliminates redundant visual tokens.

5 CONCLUSION AND FUTURE WORK

In this study, we propose an online, single-pass geometry-aware token pruning method that enhances 3D question answering performance while reducing visual token usage. Our approach projects image pixels into a shared voxel space to identify overlapped regions and generates a pruning mask to remove redundant tokens before feeding into the large language model, allowing the model to focus on novel and informative regions in 3D scenes.

Extensive experiments demonstrate that our method outperforms the online uniform sampling baseline on ScanQA, SQA3D, and OpenEQA-HM3D. Compared with the offline maximum coverage strategy, our method reduces token usage from 29.8M to 22.5M on Qwen3-VL-8B for ScanQA while improving both Exact Match and CIDEr scores, providing an efficient and effective solution for scalable 3D scene understanding in VLMs.

While our framework improves performance and reduces token usage, some limitations remain. Pruned tokens from overlapped regions may still contain useful spatial cues, and future work could explore integrating this information or combining geometry-aware pruning with token compression to further enhance efficiency. Additionally, our method is currently evaluated only on predefined datasets, and future work will explore applying our approach within embodied AI systems to handle real-time, real-world scenarios.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Xiaohan Zhang Pranav Putta Sriram Yenamandra Mikael Henaff Sneha Silwal Paul Mcvay Oleksandr Maksymets Sergio Arnaud Karmesh Yadav Qiyang Li Ben Newman Mohit Sharma Vincent Berges Shiqi Zhang Pulkit Agrawal Yonatan Bisk Dhruv Batra Mrinal Kalakrishnan Franziska Meier Chris Paxton Sasha Sax Aravind Rajeswaran Arjun Majumdar, Anurag Ajay. OpenEQA: Embodied Question Answering in the Era of Foundation Models. In *CVPR*, 2024.

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 (CVPR)*, 2022.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL `https://openreview.net/forum?id=tjZjv_qh_CE`.

Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pp. 202–221. Springer, 2020.

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 (CVPR)*, pp. 26428–26438, June 2024a.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

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.

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, and Trishul Chilimbi. M-llm based video frame selection for efficient video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13702–13712, June 2025.

Hsiang-Wei Huang, Fu-Chen Chen, Wenhao Chai, Che-Chun Su, Lu Xia, Sanghun Jung, Cheng-Yen Yang, Jenq-Neng Hwang, Min Sun, and Cheng-Hao Kuo. Zero-shot 3d question answering via voxel-based dynamic token compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19424–19434, June 2025a.

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. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

Ting Huang, Zeyu Zhang, and Hao Tang. 3d-r1: Enhancing reasoning in 3d vlms for unified scene understanding, 2025b. URL https://arxiv.org/abs/2507.23478.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984, 2024.

Benlin Liu, Yuhao Dong, Yiqin Wang, Zixian Ma, Yansong Tang, Luming Tang, Yongming Rao, Wei-Chiu Ma, and Ranjay Krishna. Coarse correspondences boost spatial-temporal reasoning in multimodal language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3783–3792, June 2025.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=IDJx97BC38.

Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. Gpt4scene: Understand 3d scenes from videos with vision-language models. *arXiv preprint arXiv:2501.01428*, 2024.

Yuan Tang, Xu Han, Xianzhi Li, Qiao Yu, Yixue Hao, Long Hu, and Min Chen. Minigpt-3d: Efficiently aligning 3d point clouds with large language models using 2d priors. *arXiv preprint arXiv:2405.01413*, 2024.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

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.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025.

Runsen Xu, Zhiwei Huang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Vlm-grounder: A vlm agent for zero-shot 3d visual grounding. In *CoRL*, 2024a.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *ECCV*, 2024b.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–22, 2023.

Jiawei Zhang, Chejian Xu, and Bo Li. Chatscene: Knowledge-enabled safety-critical scenario generation for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15459–15469, 2024.

Yiming Zhang, ZeMing Gong, and Angel X. Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15225–15236, October 2023.

Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2412.00493. arXiv preprint arXiv:2412.00493.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d capabilities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4295–4305, October 2025.

A Appendix

In this supplementary material, we provide more details in the following sections:

- In Section B, we present the categorical-level performance comparison between our method and the uniform sampling baseline, using the OpenEQA-HM3D dataset (Arjun Majumdar, 2024). The results in this experiment further corroborates the benefits of our solution by archiving superior performance while using comparable amount of tokens.
- In Section C, we report categorical-level results on VSI-Bench yang2025thinking, a benchmark designed to evaluate visual-spatial intelligence, including spatial size estimation and spatial relationship reasoning. The experimental results demonstrate that our method generalizes beyond 3D question answering benchmarks such as ScanQA and SQA3D, and remains effective on visual-spatial intelligence benchmarks.
- In Section D, we provide more details on the pruning threshold τ_o by comparing results with differentnumbers of frames as the baseline and report categorical-level performance on the SQA3D dataset (Ma et al., 2023).
- In Section E, we present more details on scaling the number of frames by reporting categorical-level performance across different model sizes and compared to uniform sampling on SQA3D.
- In Section F, we provide additional qualitative results on OpenEQA-HM3D and SQA3D.

B Categorical results on OpenEQA-HM3D.

Table 5: **Categorical results on OpenEQA-HM3D.** Comparisons on online sampling strategies, with evaluation results reported across seven categories of OpenEQA-HM3D.

Model	Sampling Strategy	OpenEQA-HM3D Arjun Majumdar (2024)								Tokens ↓
		Attr. Recog.	Func. Reason.	Obj. Local.	Obj. Recog.	Obj. State Recog.	Spatial Under.	World Know.	LLM Match ↑	
Qwen2.5-VL-7B	Uniform	63.2	61.4	51.9	49.3	48.6	45.1	**53.1**	53.1	**4.4M**
	Uniform + Ours	**71.6**	**61.8**	**54.4**	**54.5**	**63.6**	**49.6**	51.2	**58.2**	5.2M
Qwen3-VL-8B	Uniform	**71.4**	66.4	70.2	64.0	66.5	**59.8**	70.0	67.1	**3.4M**
	Uniform + Ours	69.6	**72.5**	**70.2**	**65.5**	**70.4**	53.8	**70.7**	**67.8**	4.1M

In Table 5, we report categorical-level results on the OpenEQA-HM3D dataset. The questions are categorized into seven types: (1) Attribute Recognition, (2) Functional Reasoning, (3) Object Localization, (4) Object Recognition, (5) Object State Recognition, (6) Spatial Understanding, and (7) World Knowledge. With comparable token usage, our method on top of uniform sampling yields better overall LLM Match scores. Our method demonstrates consistent improvements across nearly all categories for both Qwen3-VL and Qwen2.5-VL models. Notably, our method achieves a 15-point

improvement in LLM Match score for the "Object State Recognition" category with Qwen2.5-VL-7B, while also showing meaningful gains in other categories such as Attribute Recognition with an 8.4-point improvement.

C CATEGORICAL RESULTS ON VSI-BENCH.

Table 6: **Categorical results on VSI-Bench.** Comparisons on online sampling strategies, with evaluation results reported across seven categories of VSI-Bench.

Model	Sampling Strategy	VSIBench (ScanNet)								Avg. ↑	Tokens ↓
		Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order		
Qwen2.5-VL-7B	Uniform	33.7	9.1	**31.4**	36.7	40.5	36.4	**25.0**	28.2	29.3	16.4M
	Uniform + Ours	**34.3**	**9.5**	31.1	**40.5**	**41.3**	**39.4**	23.4	**30.3**	**30.5**	**14.8M**
Qwen3-VL-8B	Uniform	66.3	47.1	69.4	53.2	54.1	46.6	32.8	52.1	54.5	12.7M
	Uniform + Ours	**73.4**	**44.8**	**70.0**	**54.8**	**57.1**	**51.9**	32.8	**61.3**	**58.1**	**11.9M**

In Table 6, we report categorical-level results on the VSI-Bench. The benchmark consists of three datasets: ScanNet (Dai et al., 2017), ScanNet++ (Yeshwanth et al., 2023), and ARKitScenes (Baruch et al., 2021). We evaluate on the ScanNet split, where questions are categorized into eight types: (1) Object Count, (2) Absolute Distance, (3) Object Size, (4) Room Size, (5) Relative Distance, (6) Relative Direction, (7) Route Plan, and (8) Appearance Order. Following the original VSI-Bench metric, the first four categories require numerical answers while the latter four are multi-choice questions. Our method applied on top of uniform sampling achieves better average scores for both models. Notably, for Qwen3-VL-8B, our approach improves the average score from 54.5 to 58.1 while reducing token usage from 12.7M to 11.9M.

D MORE DETAILS ON THE PRUNING THRESHOLD.

Table 7: **More details on the pruning threshold τ_o.** We report six categorical results on SQA3D to investigate how different threshold settings affect pruning behavior, with higher τ_o values leading to less aggressive token removal. The results show that a threshold of 100% performs the best across different baselines. "F" denotes the number of frames sampled.

Model	Sampling Strategy	τ_o	SQA3D Ma et al. (2023)							
			what	which	can	is	how	others	Average ↑	Tokens ↓
Qwen3-VL-8B	Uniform 70F	-	46.6	45.9	54.7	62.9	47.4	51.9	51.4	74.5M
	Uniform 70F + Ours	25%	45.9	43.6	52.7	61.2	43.9	50.5	49.6	**10.7M**
	Uniform 70F + Ours	50%	46.8	43.9	52.7	63.5	45.8	51.8	50.8	13.5M
	Uniform 70F + Ours	80%	46.6	44.7	**57.7**	61.0	45.2	51.6	50.8	19.4M
	Uniform 70F + Ours	100%	**48.3**	**47.0**	56.2	**70.0**	**49.0**	**52.1**	**52.2**	35.2M
	Uniform 20F	-	45.6	45.6	54.1	60.6	46.2	**51.1**	50.2	21.6M
	Uniform 20F + Ours	25%	44.6	**46.4**	55.9	60.4	43.2	48.9	49.3	**8.3M**
	Uniform 20F + Ours	50%	44.6	43.0	55.3	61.2	43.9	48.9	49.1	9.2M
	Uniform 20F + Ours	80%	45.4	43.6	55.0	61.2	46.7	49.3	49.9	11.0M
	Uniform 20F + Ours	100%	**46.2**	43.6	**58.3**	**61.5**	**46.9**	48.8	**50.5**	14.6M

In Table 7, we further investigate the effect of the pruning threshold τ_o. From the results, we observe that applying our method on top of uniform sampling frames can significantly reduce token usage across different threshold settings. When τ_o is set to 25%, the method aggressively reduces redundant information but incurs a performance drop of 1.8 points on the 70-frame baseline. This trend

is consistent across both 70-frame and 20-frame baselines. **However, when our geometry-aware token pruning employs a milder strategy with τ_o set to 100%, it removes only highly redundant information, achieving both token reduction and performance improvement**. Specifically, with τ_o set to 100%, we achieve over 50% token reduction on the 70-frame baseline and more than 30% on the 20-frame baseline. The larger token reduction on the 70-frame baseline is due to denser frame sampling, which leads to more overlapping regions that can be pruned. This trend holds consistently across different uniform sampling frame baselines. Notably, applying our method with τ_o set to 100% on the 70-frame baseline, the "is" category achieves a more than 7-point improvement in Exact Match score.

E More Details on Scaling Sampled Frames.

Table 8: **More details on scaling the number of frames.** We report six categorical results on SQA3D using different frame counts as baselines and apply our method with different model sizes. By adopting our method, nearly all baselines achieve performance gains while using fewer tokens. "F" denotes the number of frames sampled.

Model	Sampling Strategy	SQA3D Ma et al. (2023)							
		what	which	can	is	how	others	Average ↑	Tokens ↓
Qwen3-VL-8B	Uniform 70F	46.6	45.9	54.7	**62.9**	47.4	51.9	51.4	74.5M
	+ Ours	**48.3**	**47.0**	**56.2**	62.0	**49.0**	**52.1**	**52.2** (+0.8)	**35.2M** (-53%)
	Uniform 50F	47.0	**47.9**	**58.6**	**62.4**	**47.3**	50.2	**51.6**	53.5M
	+ Ours	**47.7**	47.0	**58.6**	61.4	46.2	**51.2**	**51.6**	**27.6M** (-45%)
	Uniform 30F	46.3	47.6	55.0	**62.3**	**45.8**	50.0	**50.8**	32.2M
	+ Ours	**46.4**	**48.2**	**55.6**	60.9	44.3	**51.4**	50.7 (-0.1)	**19.4M** (-40%)
	Uniform 20F	45.6	**45.6**	54.1	60.6	46.2	**51.1**	50.2	21.6M
	+ Ours	**46.2**	43.6	**58.3**	**61.5**	**46.9**	48.8	**50.4** (+0.2)	**14.6M** (-32%)
Qwen3-VL-4B	Uniform 70F	45.5	**46.4**	60.4	60.7	**52.3**	**48.8**	51.3	74.5M
	+ Ours	**45.8**	45.3	**62.4**	**62.0**	51.2	47.7	**51.4** (+0.1)	**35.2M** (-53%)
	Uniform 50F	45.4	**47.0**	59.8	**62.3**	48.4	**48.8**	51.0	53.5M
	+ Ours	**46.6**	46.4	**60.4**	61.7	**48.8**	47.9	**51.2** (+0.2)	**27.6M** (-45%)
	Uniform 30F	44.0	47.3	62.4	60.9	48.6	**50.9**	51.0	32.2M
	+ Ours	**45.0**	**48.4**	**63.6**	**62.3**	**49.9**	50.0	**51.8** (+0.8)	**19.4M** (-40%)
	Uniform 20F	42.8	**47.3**	**61.5**	61.2	46.5	48.8	49.9	21.6M
	+ Ours	**43.9**	46.2	61.0	**61.5**	**46.9**	**50.0**	**50.4** (+0.5)	**14.6M** (-32%)

Applying our geometry-aware pruning method on top of different uniform sampling strategies with models of different sizes, Qwen3-VL-8B and Qwen3-VL-4B, we report categorical-level performance on SQA3D in Table 8. We adopt increasing uniform sampling frame counts of 20, 30, 50, and 70 as baselines and apply our method on top of each. We observe that our method consistently achieves larger performance improvements than the baselines while using fewer tokens. For Qwen3-VL-8B, the baseline achieves an Exact Match of 51.4 with 70 uniformly sampled frames. Applying our method on the same 70 frames improves the Exact Match to 52.2 while consuming less than 50% of the tokens. The same trend appears in the smaller model. From the categorical-level results, we observe that nearly all categories benefit from our method while using significantly fewer computational resources.

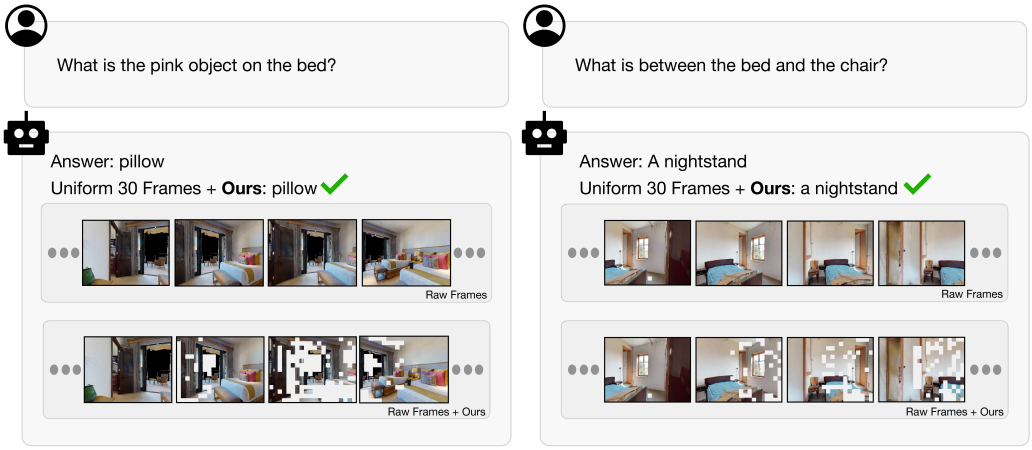

Figure 5: **Qualitative results on OpenEQA-HM3D.** We adopt our method with a uniform sampling strategy on OpenEQA-HM3D. The left and right examples show different scenes with different questions. Our method can effectively prune the overlapping regions.

F Additional Qualitative Results.

We present additional qualitative results in Figure 5 and Figure 6. The top row shows the raw frames, and the bottom row visualizes the results after applying our pruning method. We use Qwen3-VL-8B as the base model for these results.

In Figure 5, we visualize two examples from the OpenEQA-HM3D dataset to illustrate how our method prunes redundant tokens and demonstrates its effectiveness. Our method removes highly overlapping information in an online manner, progressively pruning more tokens as additional frames are received because information may be repeated across frames. In the left example, we observe that in the third image, more regions are pruned as the same visual information has already appeared in previous frames. This pruning pattern is evident across both examples.

In Figure 6, we provide six additional qualitative results from the SQA3D dataset, applying our method on top of uniformly sampled 30 frames. These examples illustrate a key limitation of uniform sampling: consecutive frames often contain highly redundant visual information. Our geometry-aware pruning method effectively removes this redundancy while preserving task-critical details. Importantly, our method maintains the model's ability to answer questions correctly while reducing token consumption and minimizing computational interference. By selectively pruning overlapping regions across frames, our method achieves a balance between computational efficiency and answer quality, which is particularly valuable in 3D question answering tasks.

Figure 6: **Additional qualitative results.** We apply our method on top of uniformly sampled 30 frames on the SQA3D dataset. The qualitative results show that our method prunes redundant information while preserving important details, thereby improving both token efficiency and performance.