

Zero-shot 3D Question Answering via Dynamic Token Compression

Hsiang-Wei Huang^{1*}, Fu-Chen Chen², Wenhao Chai¹, Che-Chun Su², Lu Xia², Sanghun Jung¹
Cheng-Yen Yang¹, Jenq-Neng Hwang¹, Min Sun², and Cheng-Hao Kuo²

University of Washington¹ Amazon²

* Work done in Amazon applied scientist internship.

Motivation

Vision Language Models (VLMs) suffer from memory and computational overhead when dealing with 3D scan video.

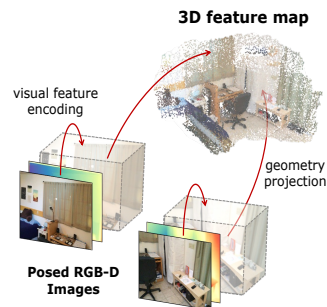
Introduction

- We proposed two token compression methods leveraging spatial prior and visual semantic to reduce the visual token and enhance VLM's memory and computational efficiency.
- Our proposed token compression can reduce the visual token by 90% while maintaining the performance.
- Comparable with SoTA performance on OpenEQA, ScanQA, and SQA3D.

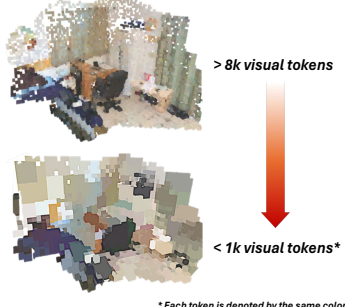
TLDR -- We propose a token compression method that achieve 90% token reduction while maintaining competitive performance on 3D question answering task.

Method Overview

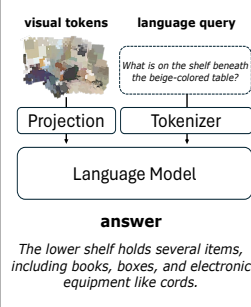
Step 1: Multi-view 2D feature encoding.



Step 2: Token compression.



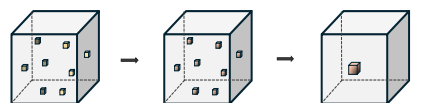
Step 3: LLM understanding.



- Step 1: Multi-view 2D feature encoding.
- Step 2: Visual tokenization.
 - simple concatenate → > 8,000 # v-tokens
 - simple 3D voxelize → > 4,000 # v-tokens
 - dynamic 3D voxelization → < 1,000 # v-tokens
- Step 3: LLM understanding and reasoning.

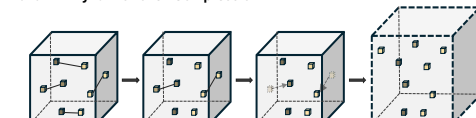
Token Compression

Variant A: Vanilla Token Compression



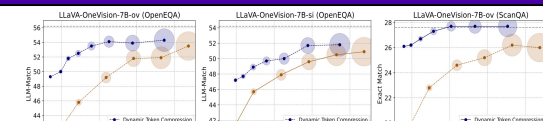
- Voxelize the 3D point cloud field. Each \square is a visual token extracted from multi-view 2D image.
- Conduct average pooling among all visual tokens within the voxel.
- Each voxel resulted in one visual token. The resulted visual tokens will be sent to LLM.

Variant B: Dynamic Token Compression



- Draw edges between visual tokens in each voxel.
- Keep the $r\%$ most similar edges.
- Compress connected token pairs.
- Compression in next-level voxel scale.

Experiments



Models	LLM-Match	Models	EM@1	Visual token usage	Sampl.	VTC	DTC
Cloud-source		Task-specific models					
LLaMA-2 70B ¹	28.3	VanNeuMCAN	17.3	100%	56.2	-	-
GPTE ¹	33.5	ScanQA	21.1	43%	51.9	54.2	54.3
Claude-3 Opus	36.3	3D-VLP (Jin et al.)	21.7	26%	49.2	50.5	54.1
Gemini 1.0 Pro Vision	44.9	3D-ViTA	22.4	17%	45.8	47.4	52.5
Claude-3.5 Sonnet	48.7	3D-VLP (Zhang et al.)	23.0	8%	40.4	43.6	49.3
GPTE-V (15 frames)	54.6	Video-LLMs					
GPTE-V (50 frames)	55.3	AzureCap	17.2				
Open-source		Agent3D Zero	17.5				
Video-LLaMA	20.0	LLaVA-NaXT-Video	18.7				
LLaMA-2 w/ Concept Graph	28.7	VideoChat2	19.2				
AzureCap	28.9	MovieChat (w/ LLaVA-OV-7B)	20.0				
Video-ChatPT	32.1	Task-specific fine-tuned 3D-LLMs					
LLaMA-2 w/ Sparse Visual Map	34.3	3D-LLM	20.5				
LLaMA-2 w/ LLaVA-1.5 caption	36.8	FD-ScanQA	22.9				
Chat-LLaVA	42.3	LEO	24.5				
Video-LLaMA2	49.2	Scene-LLM	27.2				
MovieChat (w/ LLaVA-OV-7B)	54.9						
Ours		Ours					
Base Model (100%)	56.2	Base Model (100%)	27.6				
w/ DTC (50%)	55.3 (+0.8)	w/ DTC (54%)	27.8 (+0.2)				
w/ DTC (43%)	54.3 (+1.9)	w/ DTC (40%)	27.7 (+0.1)				
w/ DTC (20%)	54.1 (+2.1)	w/ DTC (12%)	27.1 (+0.1)				
w/ DTC (17%)	52.5 (+3.7)	w/ DTC (14%)	26.7 (+0.9)				
w/ DTC (8%)	49.3 (+6.9)	w/ DTC (9%)	26.1 (+1.5)				

Qualitative Results



3D Question Answering using less than 1,000 v-tokens per scene!