HALL A
Thu, Jul 9, 2026 • 5:00 PM – 6:45 PM KST

CG-MLLM: Captioning and Generating 3D Content via Multi-modal Large Language Models

Junming Huang[1 2] Chi Wang[1] Letian Li[1 2] Guangkai Xu[1]
Donglin Huang[1] Hao Chen[1] Qiang Dai[2] Weiwei Xu[1]

[1] Zhejiang University

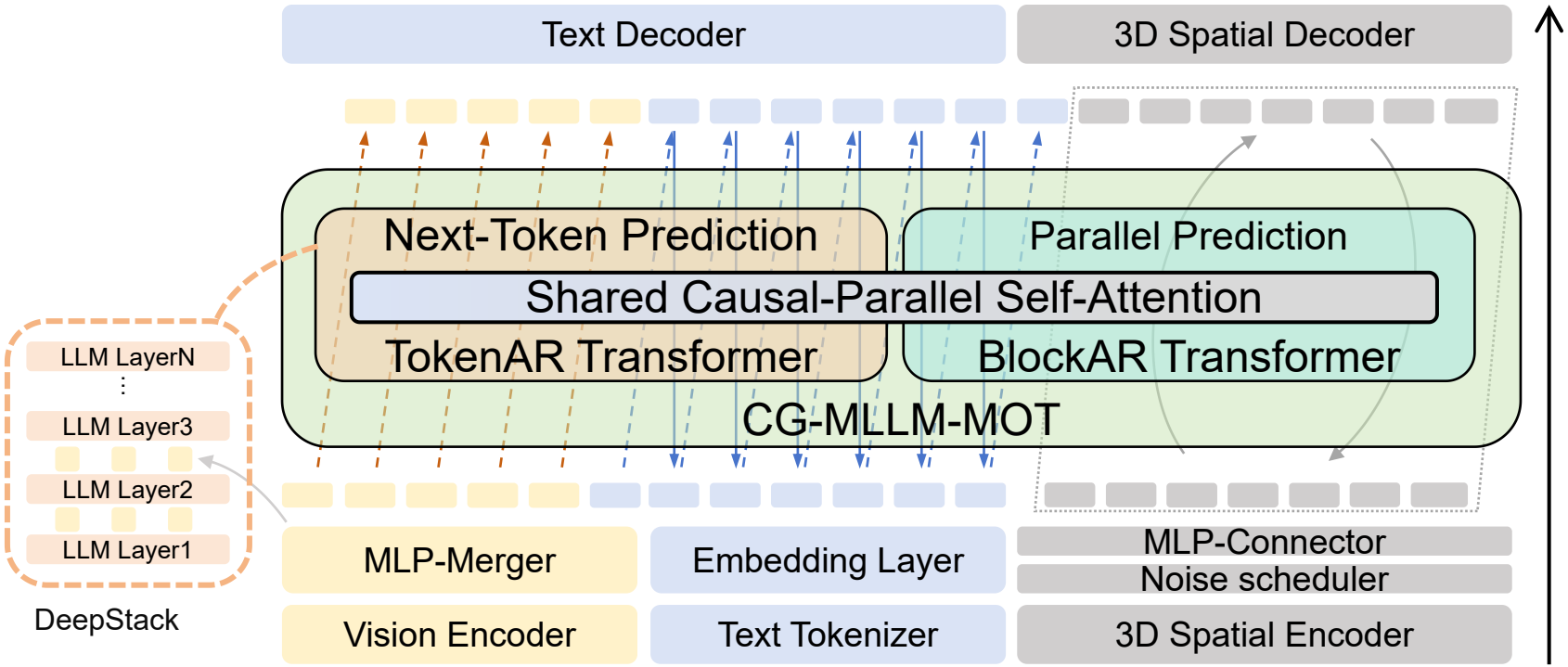
[2] LIGHTSPEED

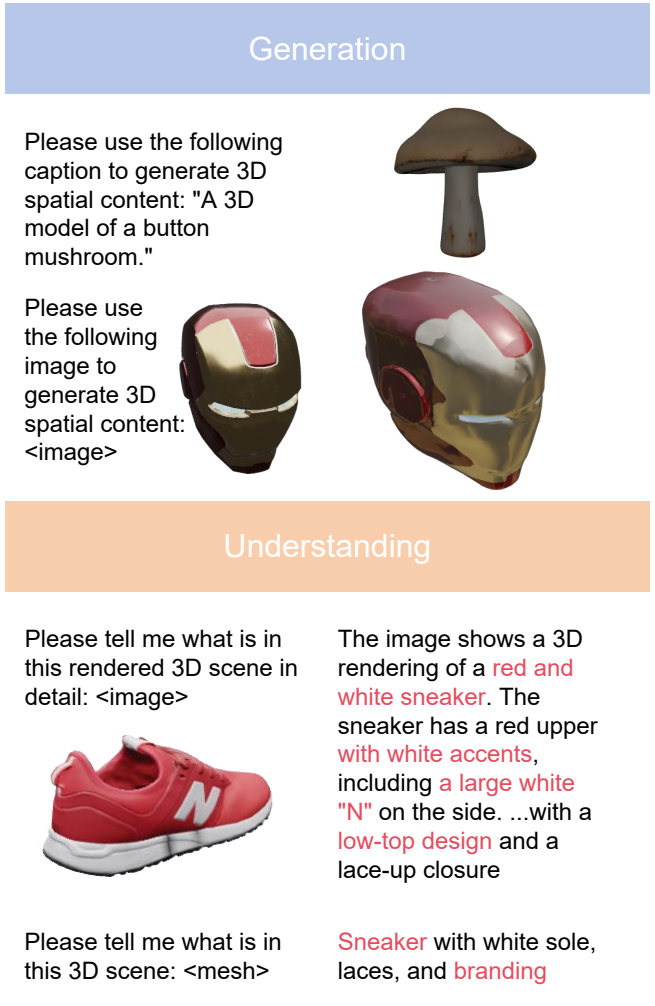
Introduction

[1] Zhejiang University
[2] LIGHTSPEED

In this paper, we propose CG-MLLM, a novel Multi-modal Large Language Model (MLLM) capable of 3D captioning and high-resolution 3D generation in a single framework. Leveraging the Mixture-of-Transformer architecture, CG-MLLM decouples disparate modeling needs, where

- **the Token-level Autoregressive (TokenAR) Transformer handles token-level content,**
- **the Block-level Autoregressive (BlockAR) Transformer handles block-level content.**

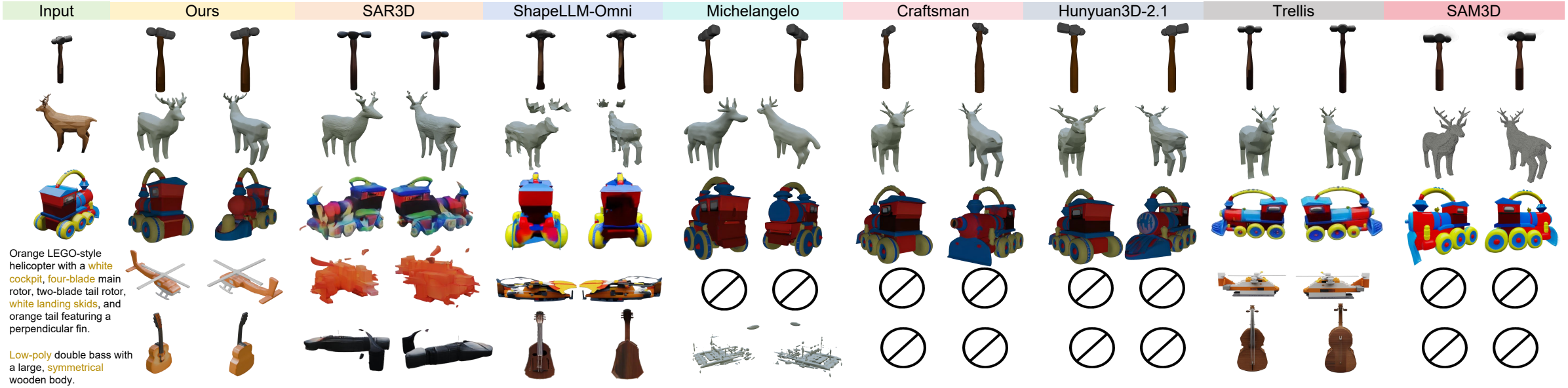




Generation Results

[1] Zhejiang University
[2] LIGHTSPEED

Both quantitative and qualitative experiments demonstrate that our approach achieves the best performance among the compared MLLM-based methods, while attaining quality comparable to non-MLLM methods.



Method	p-FID↓	p-KID↓	CLIP-IQA+↑	MUSIQ↑	CLIP↑	User Study↑
Non-MLLM-Based						
Michelangelo	17.96	0.56	0.45	71.42	84.08	2.60
CraftsMan	14.09	0.40	0.45	71.09	84.86	3.15
Hunyuan3D-2.1	16.80	0.53	0.47	71.20	85.11	3.15
TRELLIS	7.36	0.12	0.44	66.97	84.13	3.28
SAM3D	33.92	1.13	0.47	70.21	84.67	3.45
MLLM-Based						
SAR3D	30.07	1.00	0.42	66.01	82.86	2.93
ShapeLLM-Omni	13.11	0.29	0.37	55.71	84.18	2.30
Ours	12.55	0.27	0.45	71.65	84.47	3.32

Understanding Results

[1] Zhejiang University
[2] LIGHTSPEED

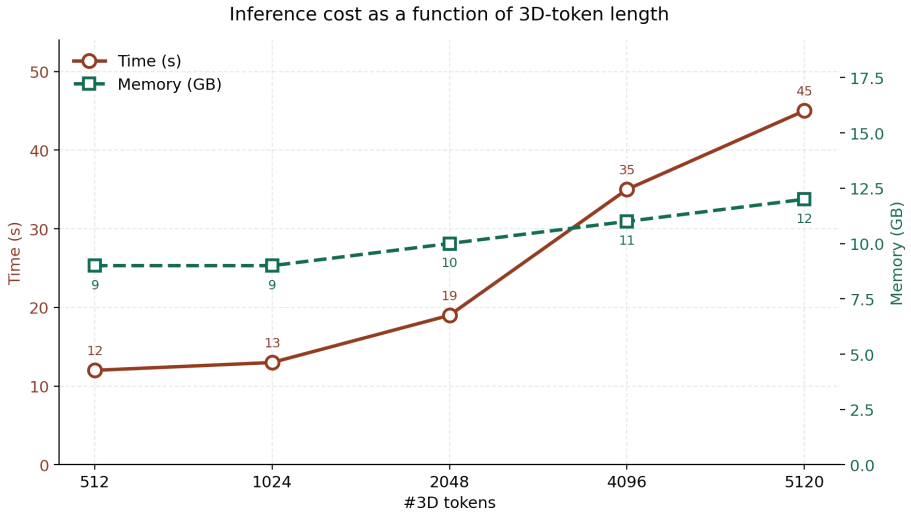
An interesting observation is that, although no dedicated object-understanding dataset was introduced on the vision side, the model's comprehension of 3D rendered images improved in tandem with generation training. (Qwen3-VL-2B vs. Ours-2B)

A plausible explanation is that, in order to "understand an image for generating its 3D counterpart," the model is implicitly forced to acquire spatial reasoning over images, **its understanding capability is bootstrapped by the generation objective**.

Model	BLEU-1↑	ROUGE-L↑	METEOR↑
3D latent Inputs			
3D-LLM	16.91	19.48	19.73
PointLLM-13B	17.09	20.99	16.45
ShapeLLM-Omni-7B	18.51	21.37	19.89
Image Inputs			
InstructBLIP-13B	4.65	8.85	13.23
LLaVA-13B	4.02	8.15	12.58
Qwen3-VL-2B	3.13	7.21	11.92
Ours-2B MOT	13.51	19.13	14.28

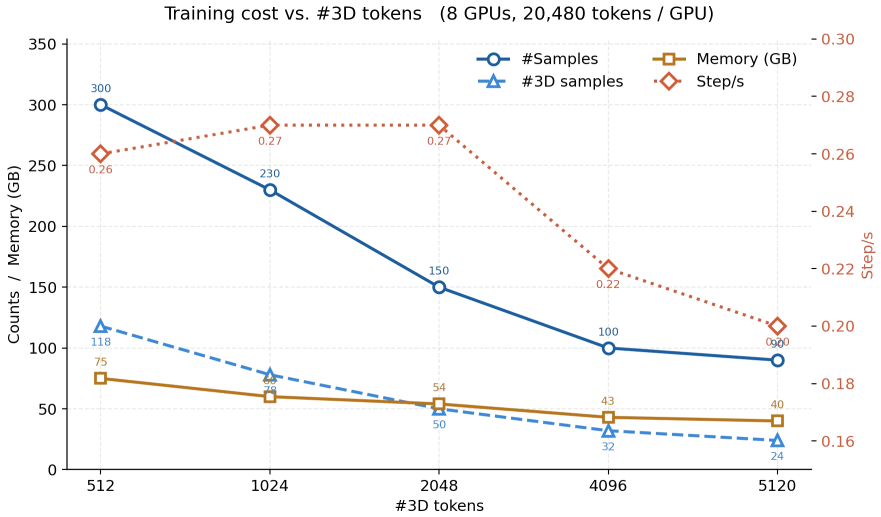
Albation Study

[1] Zhejiang University
[2] LIGHTSPEED



Ablation Study provide further insights and a clearer understanding of our model. For more details,

Scan the QR code.



HY2.1-VAE	MOT	LLM Backbone	#Tokens	p-FID ↓	p-KID ↓
✗	✗	Qwen2.5-0.5B	512	53.66	1.76
✓	✗	Qwen2.5-0.5B	512	44.91	1.42
✓	✓	Qwen2.5-0.5B	512	30.60	0.77
✓	✓	Qwen3VL-2B	512	15.61	0.43
✓	✓	Qwen2.5-0.5B	4096	16.57	0.53
✓	✓	Qwen3VL-2B	4096	**12.55**	**0.27**