

CG-MLLM

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

ICML 2026

Junming Huang, Chi Wang, Letian Li, Guangkai Xu,
Donglin Huang, Hao Chen, Qiang Dai, Weiwei Xu*



浙江大學
ZHEJIANG UNIVERSITY



LIGHTSPEED
STUDIOS

研究动机

Motivation

MLLM

文本模态过于抽象

MLLM 扩展并引入图像等多模态，增强现实感知能力

长处

强语言语义理解与多模态对齐
同时享有 LLM Scaling 红利

短板

大多仍停留于平面视觉
立体空间感知能力较弱

3D 生成与理解

随着 Objaverse、TRELLIS-500K 等数据集涌现，原生处理 3D 内容成为现实

现有 3D 生成模型（非 LLM）Scale-up 收益递减

现有 3D-MLLM 扩展受限于设计

- ① low poly 表征：抽象、精度受限
- ② 粗粒度代理：仍需额外 3D 生成模型

win-win 方案：结合 MLLM 与 3D 生成与理解

增强 MLLM 的立体空间感知，同时让 3D 理解与生成借助 MLLM 社区强大的预训练模型能力

研究缺口 目前没有兼具「3D 原理解 + 生成」能力，并能承接 LLM Scaling 趋势的模型

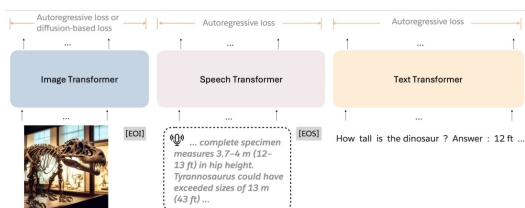
在此背景下我们提出 CG-MLLM

语言-图像-3D 原生一体化，单一 MLLM 框架端到端理解与生成

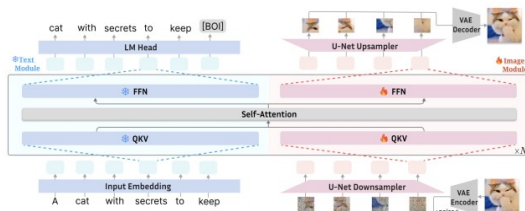
现有工作

Existing Work

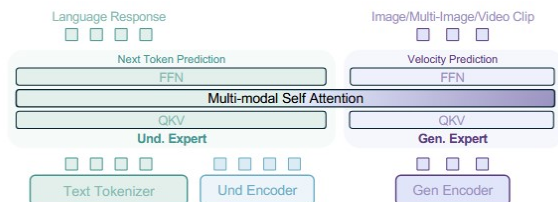
MLLM GEN / UND



Mixture-of-Transformers: A Sparse and Scalable Architecture for Multi-Modal Foundation Models

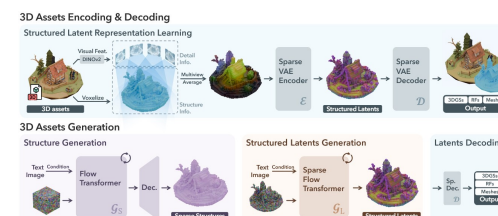


LMFusion: Adapting Pretrained Language Models for Multimodal Generation

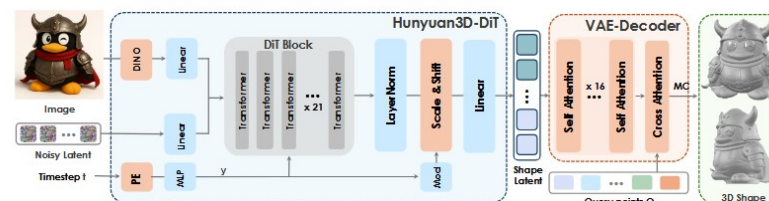


Emerging Properties in Unified Multimodal Pretraining

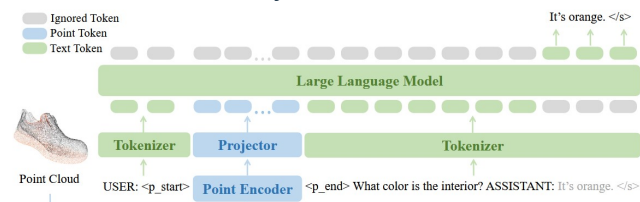
3D GEN / UND



Structured 3D Latents for Scalable and Versatile 3D Generation



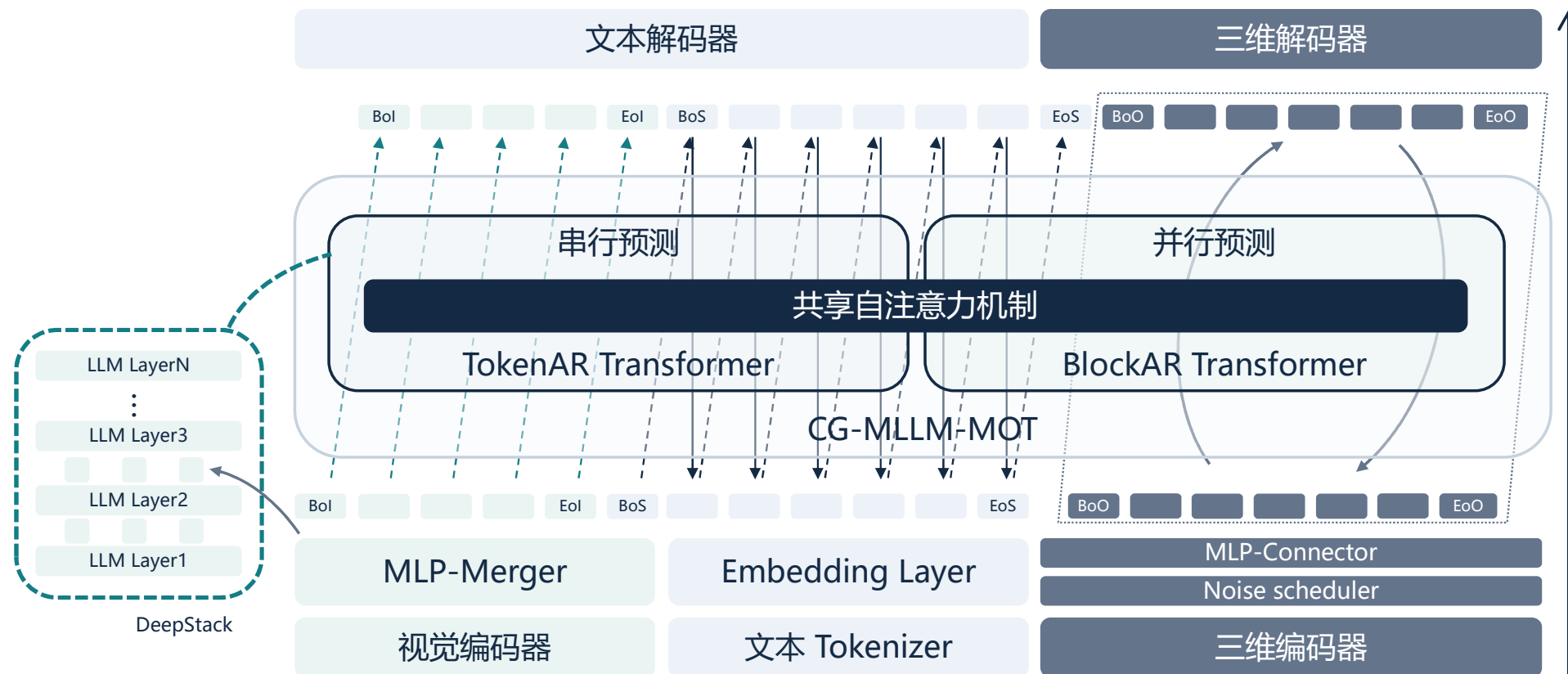
Hunyuan3D 2.1: From Images to High-Fidelity 3D Assets with Production-Ready PBR Material



PointLLM: Empowering Large Language Models to Understand Point Clouds

方法主线

Pipeline



核心思路：受 Mixture-of-Transformers 启发，为 Token 级串行建模（TokenAR）与 Block 级并行建模（BlockAR）分别配置专用 Transformer，共享因果-并行自注意力，在同一架构内融合序列语言意图与并行 3D 空间构建。

混合感受野

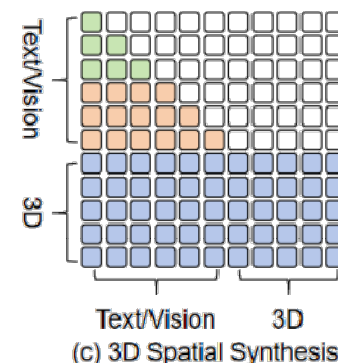
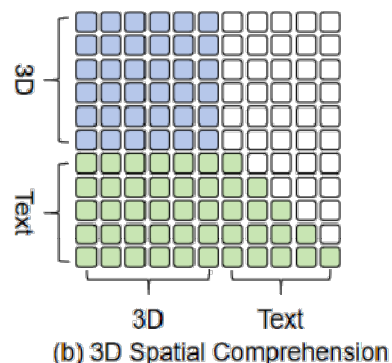
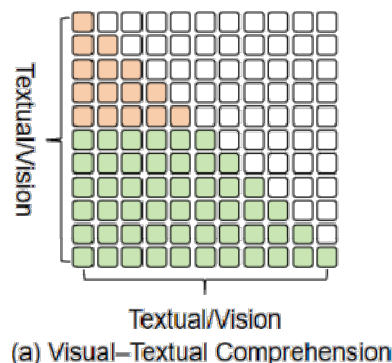
Mixed Receptive Field

因果掩码

标准自回归模型仅用因果掩码
只看前文

并行掩码

扩散模型通常使用并行掩码
样本内全可见



三种典型模式

(a) 视觉-文本理解 (b) 3D 空间理解 (c) 3D 空间合成

CG-MLLM 受先驱工作启发，自适应融合两者，构建混合 Causal-Parallel 掩码

使编码器可按其原生设计方式灵活接入对应 Transformer

3 倍

在 4096 Tokens 的设定下，通过混合感受野并行 inference，
相较于纯串行推理提升 3 倍

训练策略

01 对齐阶段

丢弃 90% 条件输入，
训练无条件生成能力与初始理解能力

3D 分辨率固定为 512 token

实现细节

Classifier-Free Guidance、logit-normal 时间步采样、AdamW
学习率 $1e-4 \rightarrow 5e-5$

训练规模

16 × NVIDIA H20 GPU，最大序列长度 36,864 → 51,200
推理 CFG scale = 7.5，采样 50 步

两个细节

AdaLN 在 MLLM 架构中反而降低稳定性
VAE 重建应仅用均匀采样点以避免孔洞畸变

02 渐进分辨率阶段

分辨率从 512 逐步提升到 4096 token

丢弃概率由 90% 降至 10%

实验结果

Result

通过文本或图片生成三维物体

Please use the following caption to generate 3D spatial content:
"A 3D model of a button mushroom."



Please use the following image to generate 3D spatial content:
<image>



通过渲染图或网格理解三维物体

Please tell me what is in this rendered 3D scene in detail: <image>



The image shows a 3D rendering of a red and white sneaker. The sneaker has a red upper with white accents, including a large white "N" on the side. ...with a low-top design and a lace-up closure

Please tell me what is in this 3D scene: <mesh>

Sneaker with white sole, laces, and branding



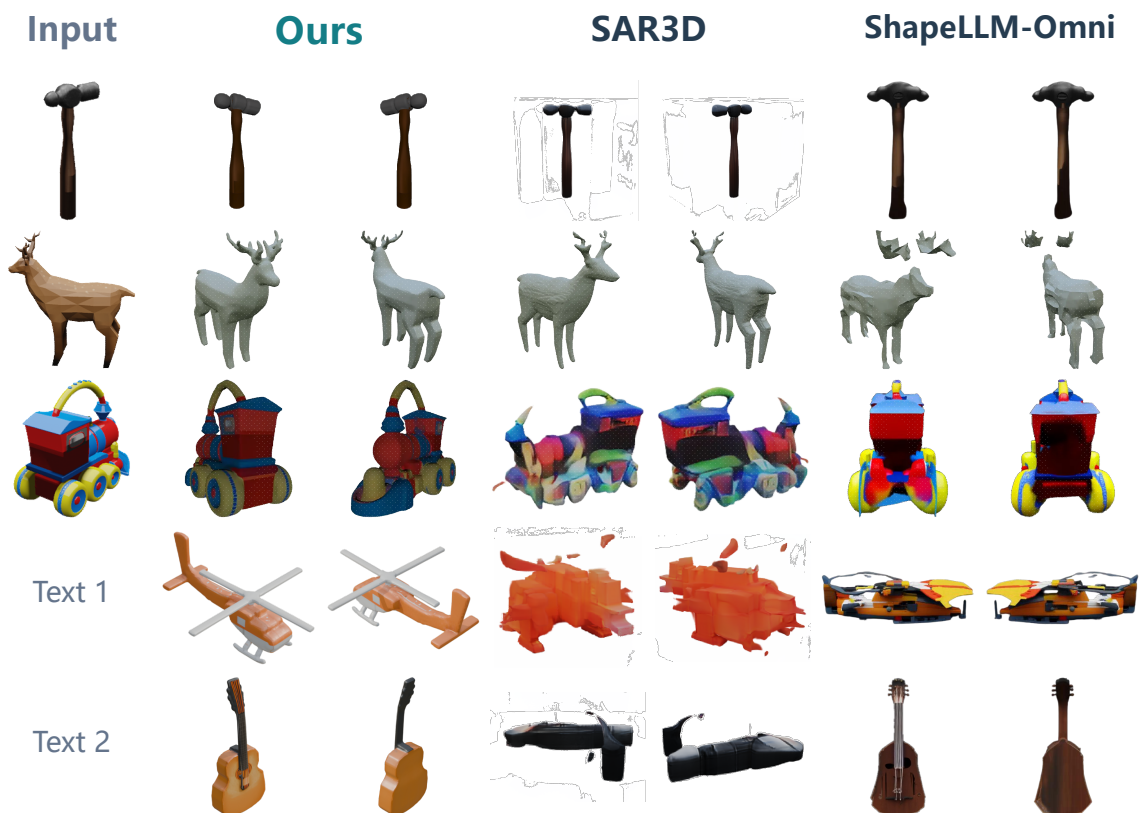
Model			
GT	A white drawing.	Maroon and black saloon car.	A cartoon old computer.
Ours	A large white sphere positioned centrally on a flat , white rectangular background. The sphere is slightly smaller in diameter than the background.	A car. It appears to be a modern, sleek vehicle with a streamlined design. The car is predominantly red , with a bright red body and black wheels .	A vintage computer setup with a small, white, rectangular computer and a matching keyboard . The computer displays a green text-based interface .

对比实验

Comparison

Method	p-FID↓	p-KID↓
Non-MLLM-Based		
Michelangelo	17.96	0.56
CraftsMan	14.09	0.40
Hunyuan3D-2.1	16.80	0.53
TRELLIS	7.36	0.12
MLLM-Based		
SAR3D	30.07	1.00
ShapeLLM-Omni	13.11	0.29
Ours	12.55	0.27

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



Text 1 Orange LEGO-style helicopter with a white cockpit, four-blade main rotor, two-blade tail rotor, white landing skids, and orange tail featuring a perpendicular fin.

Text 2 Low-poly double bass with a large, symmetrical wooden body.

结论

Summarize

- 01 CG-MLLM 提出了首个支持端到端 3D 空间生成的多模态大语言模型
- 02 证明 MLLM 无需依赖外部生成器即可原生完成精细 3D 生成
- 03 在 3D 理解与基于 LLM 的 3D 生成任务上均取得业界领先性能
- 04 揭示生成能力与感知能力可相互强化的新现象

未来方向

引入更强的多编码器图像条件机制

扩大模型参数规模与 3D token 预算

迈向完全统一的多模态生成与理解模型