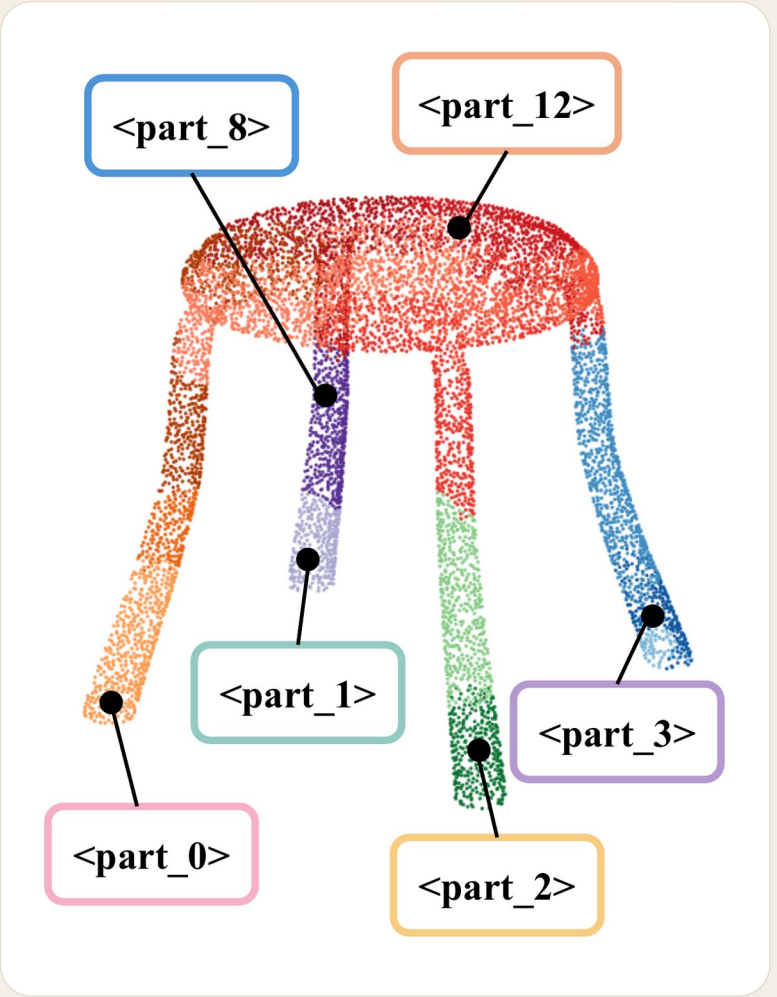
NEURIPS 2026

3D-PLOT-LLM

Part-Level Object Tokens for 3D Large Language Models

Jintang Xue[1] · Xinyu Wang[1] · Yixing Wu[1]
Jingwen Chen[2] · C.-C. Jay Kuo[1]

[1]University of Southern California · [2]Ohio State University

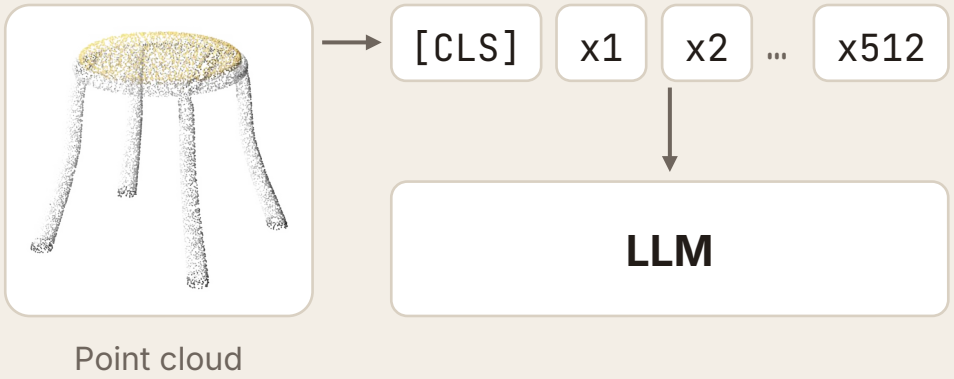


PROBLEM

3D MLLMs see objects whole, not in parts

Object-level 3D MLLMs

PointLLM and similar models feed the object to the LLM as one flat stream of patch tokens.



Point cloud

- ✓ Can caption the whole object
- ⚠ Has no token for “the legs” or “the back”

Prior part-aware 3D MLLMs

They add dense-prediction machinery around the LLM:

Kestrel

LLM → [SEG] → **Seg. decoder** → part mask

Part-X-MLLM

Dual-pathway encoder → LLM → box tokens

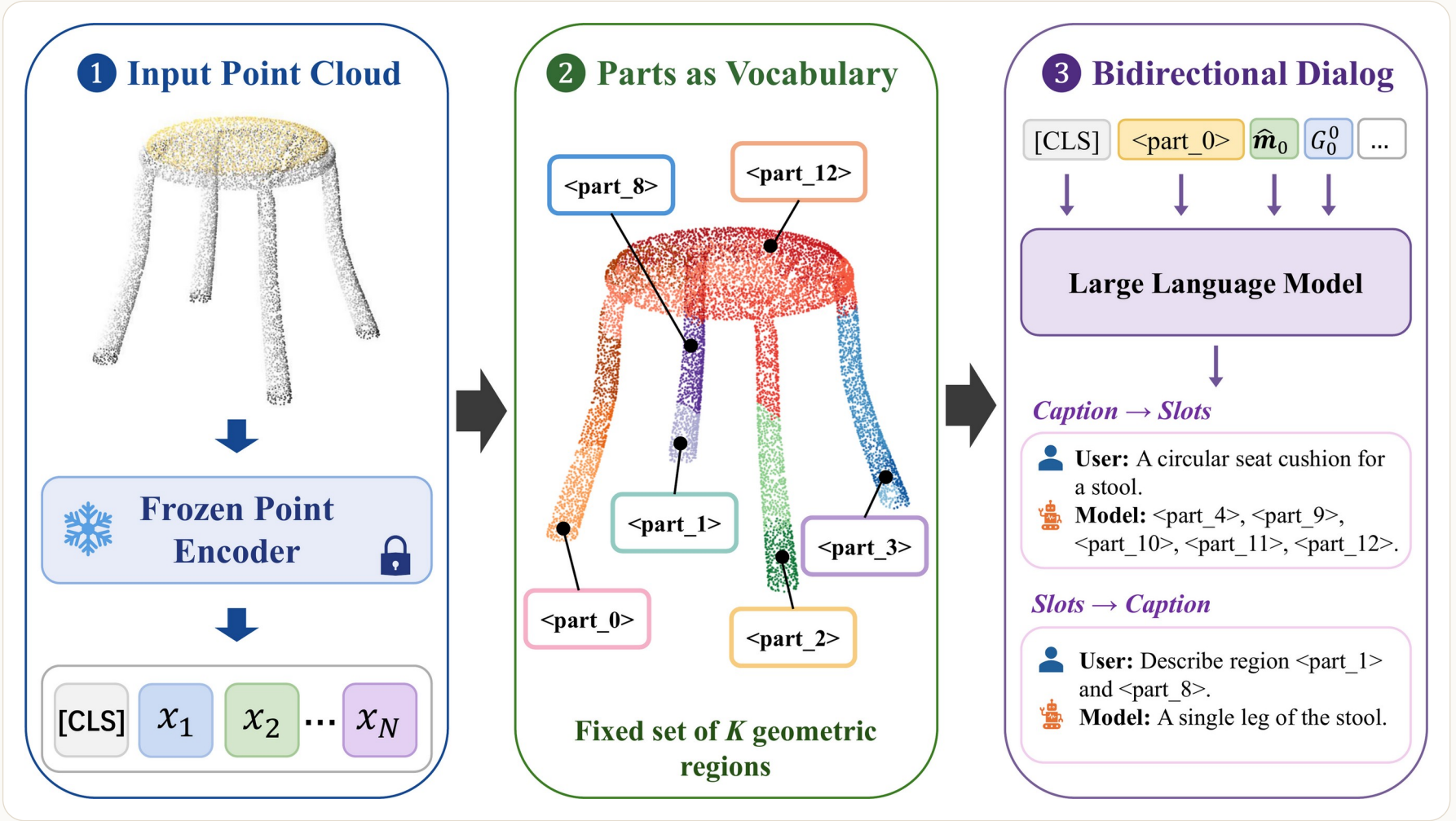
▭ added machinery outside the LLM

~10× more non-LLM parameters, and parts stay outside the LLM’s vocabulary

Can the LLM read and write parts with its own tokens?

IDEA

Each object region becomes a vocabulary token



K = 16 regions

fixed and training-free, built from the frozen encoder's patches

One token each

<part_k> joins the LLM's own vocabulary

Both directions

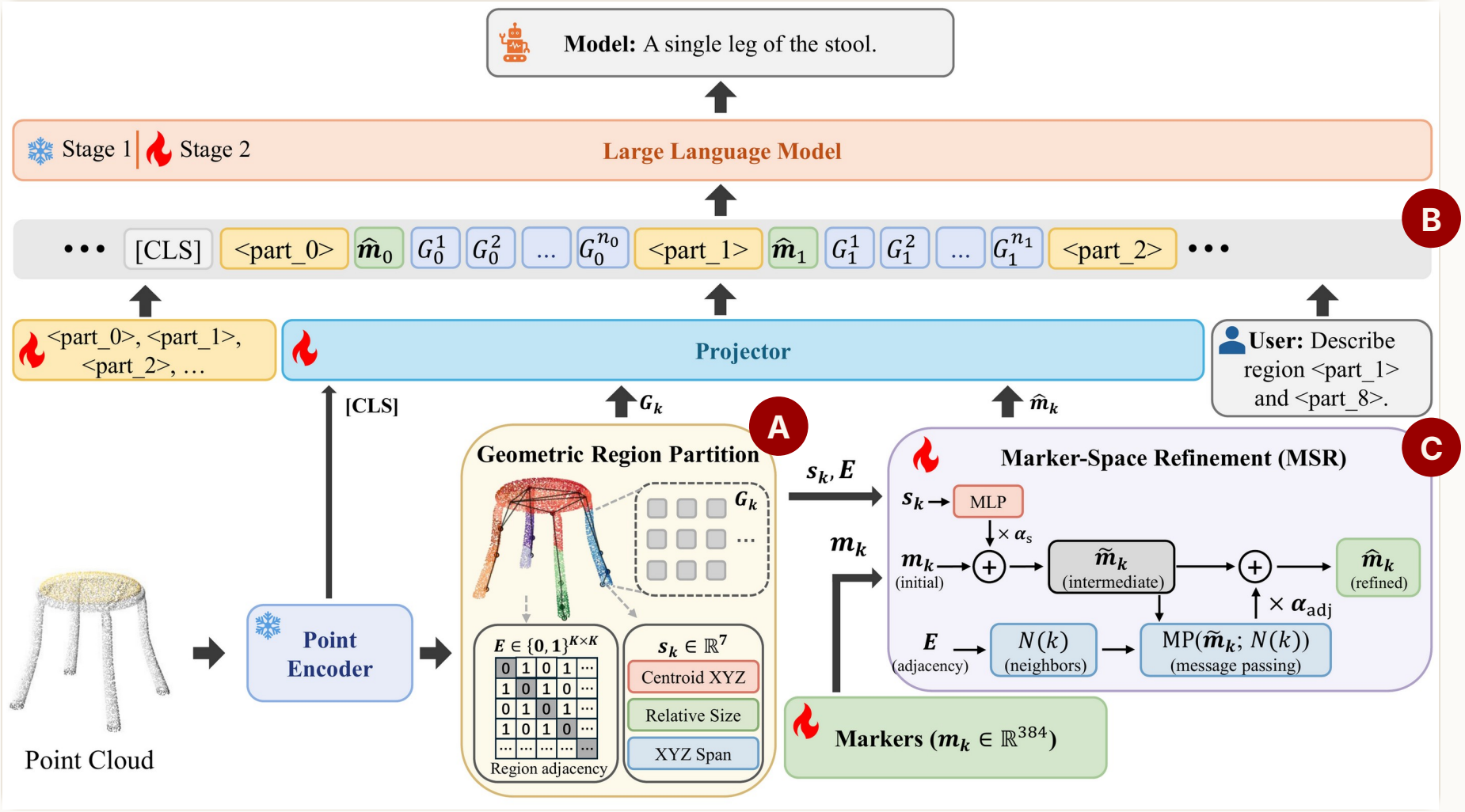
C2S caption → slots

S2C slots → caption

A named part is a set of region tokens, learned from data.

METHOD

Three steps run before the LLM



A Partition

512 frozen patches grouped into 16 regions, deterministic and training-free; each keeps its centroid, size, span, and neighbors

B Assembly

<part_k> and a learnable marker go before each region's patches

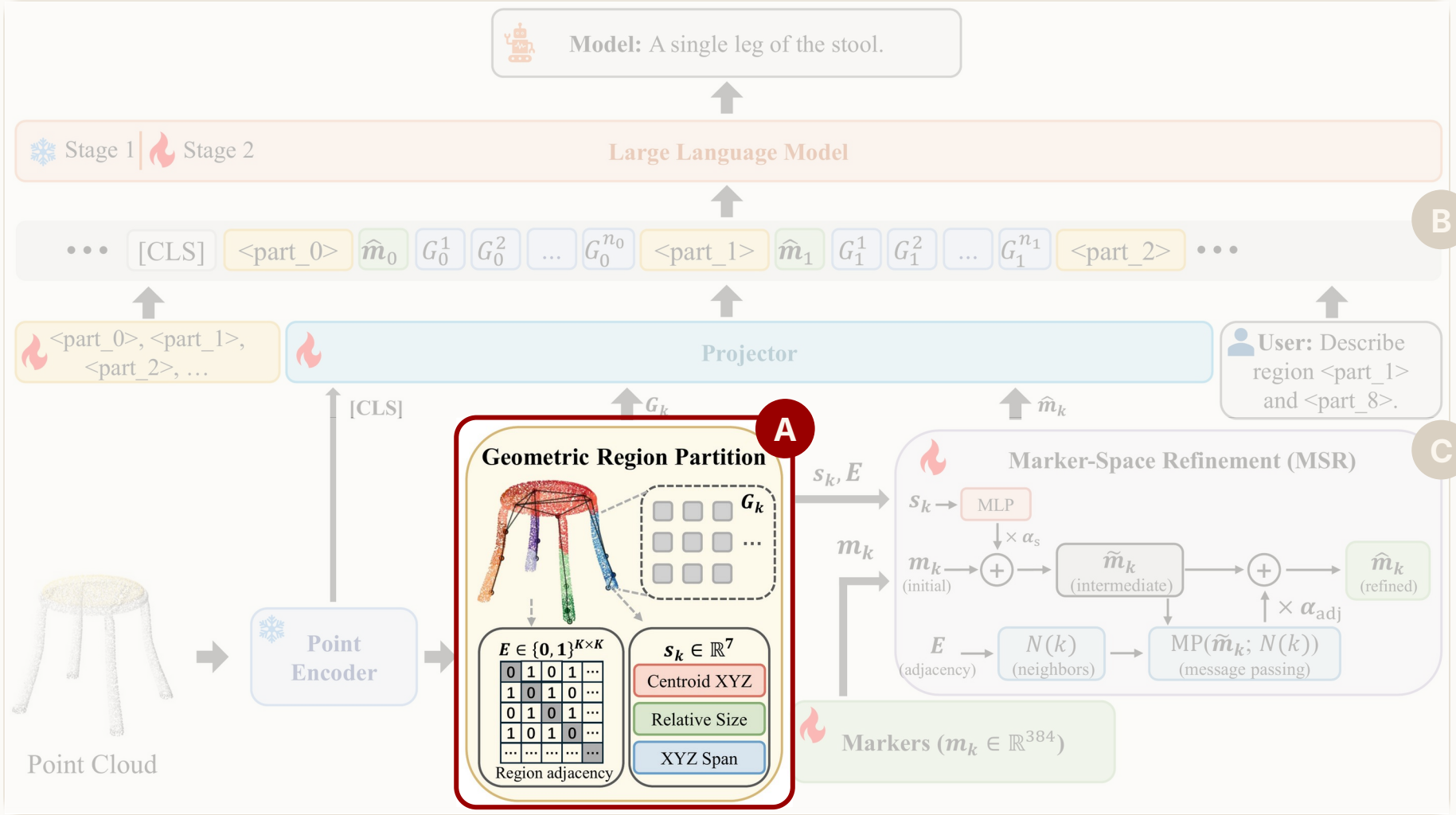
C MSR

two residuals add the region's statistics and its neighbors to each marker

≈0.8M new parameters | Frozen point encoder | Stage 1: 660K Cap3D captions | Stage 2: 70K PointLLM + 77K PartVerse-QA

METHOD

Three steps run before the LLM



A Partition

512 frozen patches grouped into 16 regions, deterministic and training-free; each keeps its centroid, size, span, and neighbors

B Assembly

<part_k> and a learnable marker go before each region's patches

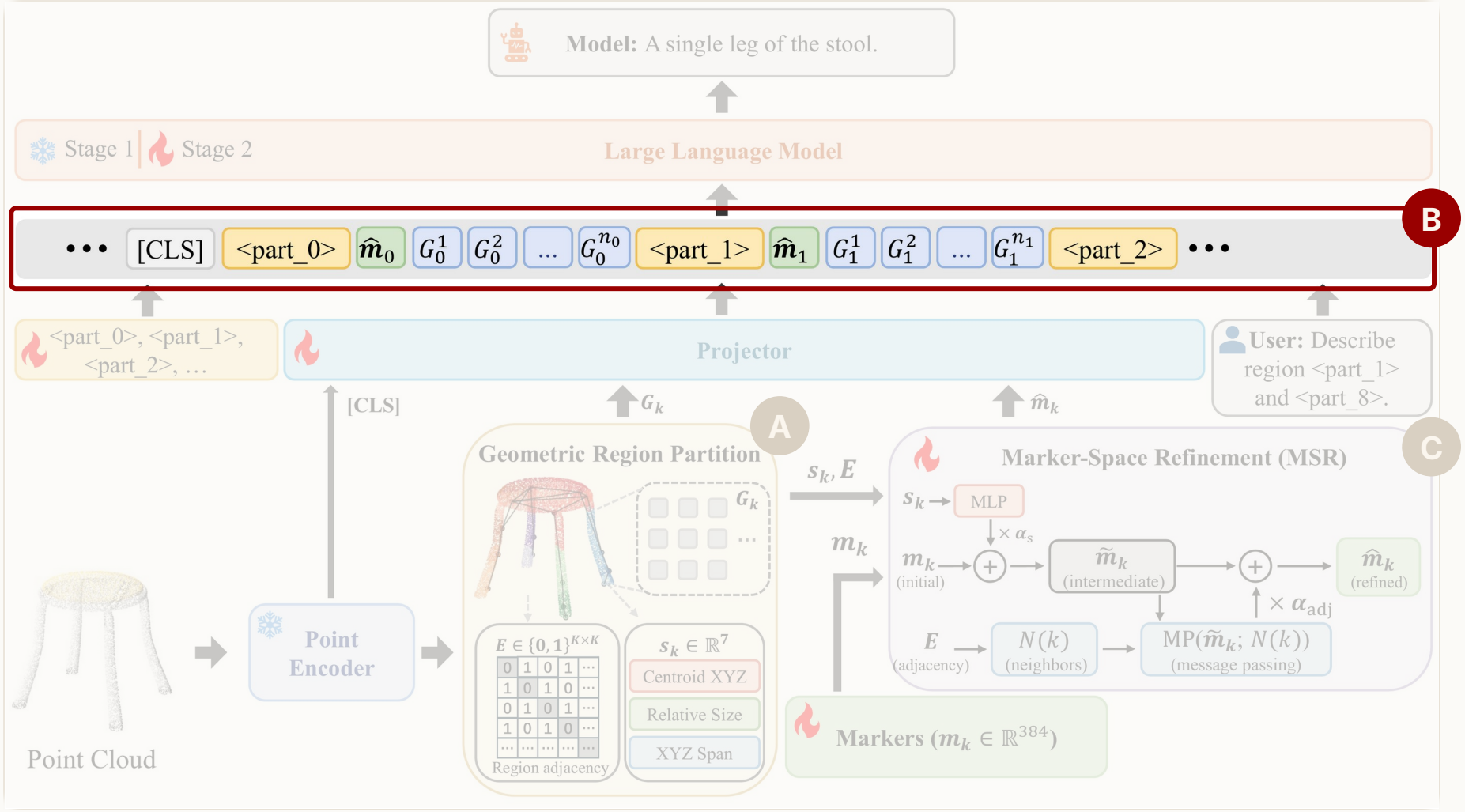
C MSR

two residuals add the region's statistics and its neighbors to each marker

≈0.8M new parameters | Frozen point encoder | Stage 1: 660K Cap3D captions | Stage 2: 70K PointLLM + 77K PartVerse-QA

METHOD

Three steps run before the LLM



A Partition

512 frozen patches grouped into 16 regions, deterministic and training-free; each keeps its centroid, size, span, and neighbors

B Assembly

<part_k> and a learnable marker go before each region's patches

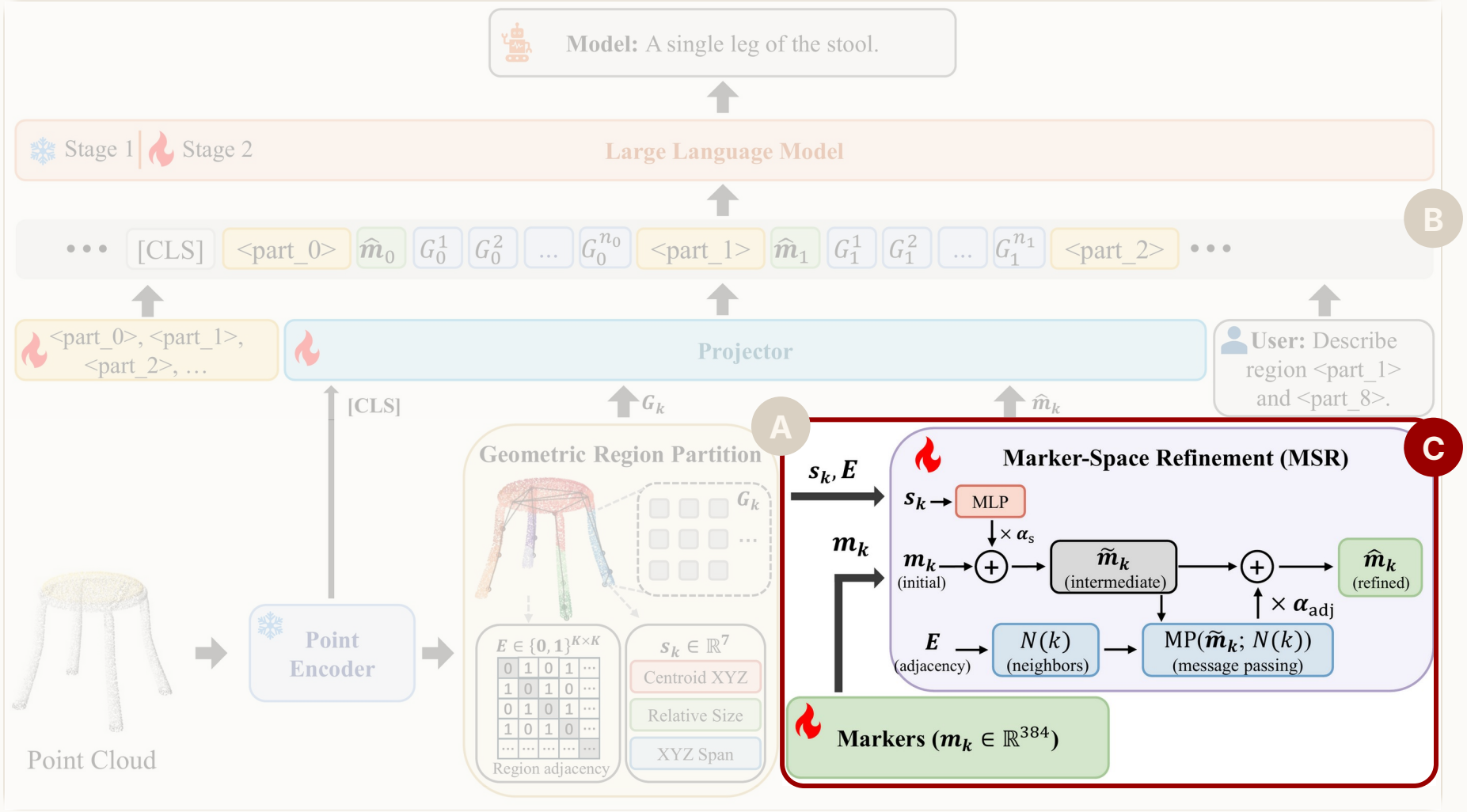
C MSR

two residuals add the region's statistics and its neighbors to each marker

≈0.8M new parameters | Frozen point encoder | Stage 1: 660K Cap3D captions | Stage 2: 70K PointLLM + 77K PartVerse-QA

METHOD

Three steps run before the LLM



A Partition

512 frozen patches grouped into 16 regions, deterministic and training-free; each keeps its centroid, size, span, and neighbors

B Assembly

<part_k> and a learnable marker go before each region's patches

C MSR

two residuals add the region's statistics and its neighbors to each marker

≈0.8M new parameters | Frozen point encoder | Stage 1: 660K Cap3D captions | Stage 2: 70K PointLLM + 77K PartVerse-QA

BENCHMARK

PartVerse-QA tests the tokens in both directions

C2S **Caption → slots**

Given a part caption, output its <part_k> set.

Scored by Jaccard and exact match

S2C **Slots → caption**

Given a <part_k> set, describe the part.

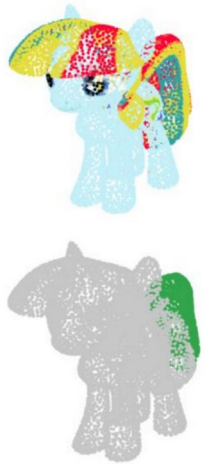
Scored by SBERT and a GPT-4o judge

77.6K training pairs

588 held-out queries
392 C2S + 196 S2C

83 held-out objects
never seen in training

(a) C2S

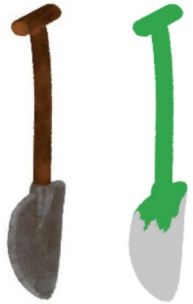


Caption:
A close-up of a pony's tail with red and yellow stripes.

GT Slots:
<part_0>, <part_5>, <part_10>, <part_12>, <part_14>.

3D-PLOT-LLM Pred:
<part_0>, <part_5>, <part_10>, <part_12>, <part_14>.

(b) S2C

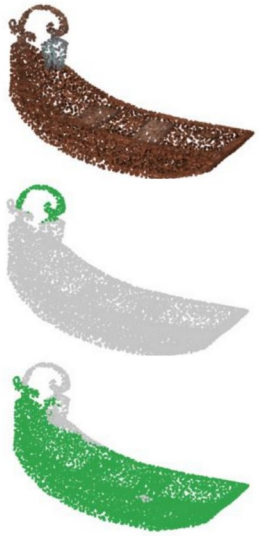


Input Slots:
<part_2>, <part_4>, <part_7>, <part_10>, <part_12>, <part_13>, <part_14>.

GT Caption:
A wooden handle of a shovel.

3D-PLOT-LLM Pred:
A wooden handle of a shovel.

(c) Bidirectional



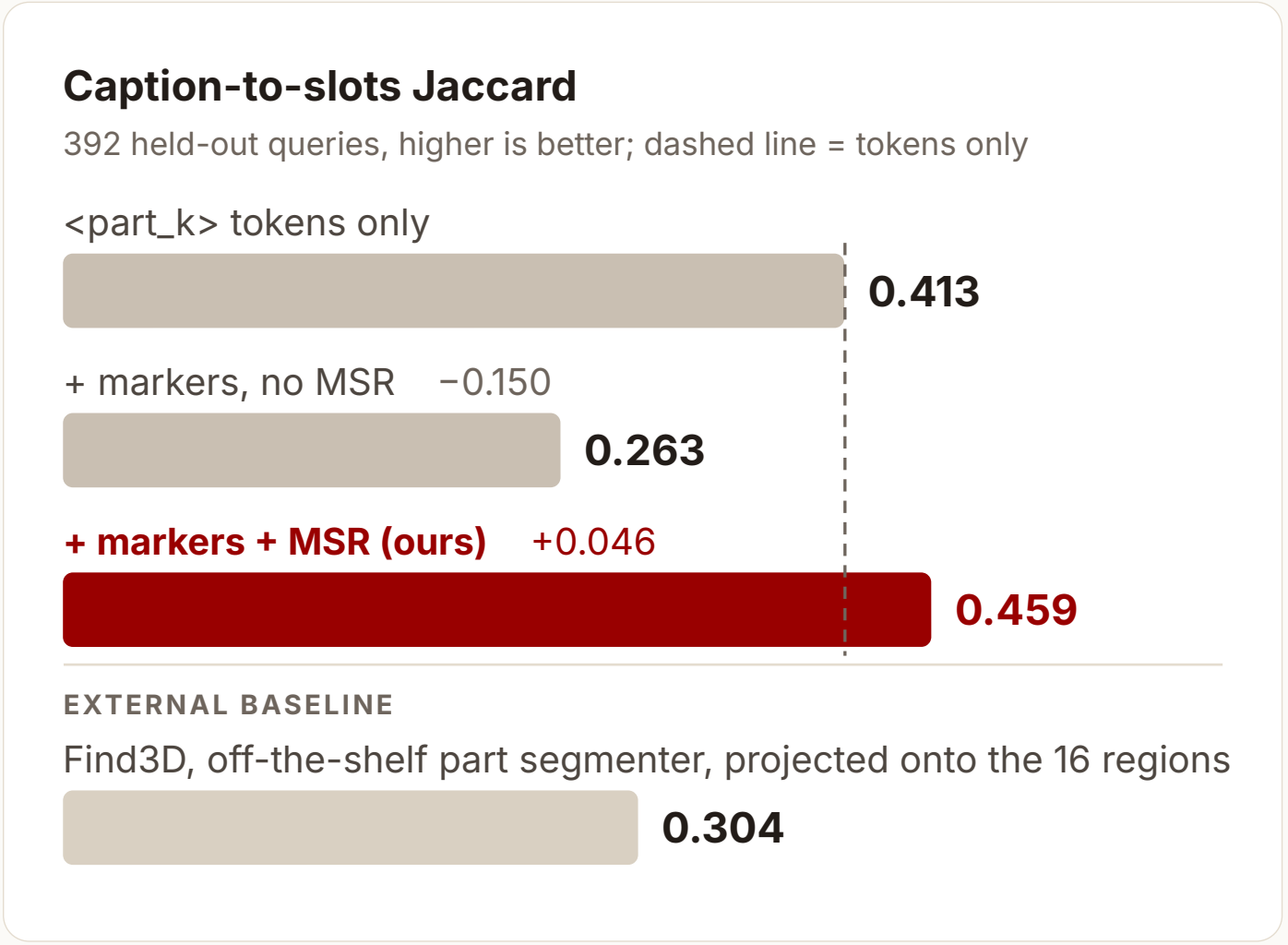
1) C2S Caption:
A black lantern component extracted from a wooden boat.
3D-PLOT-LLM Pred / GT:
<part_0> || <part_0>.

2) S2C Input:
<part_1>, …, <part_15>.
GT Caption:
A wooden boat component, specifically the hull section.
3D-PLOT-LLM Pred:
A wooden boat component, specifically the hull section.

Selected examples on held-out objects, from the paper

RESULTS · PART ADDRESSING

MSR turns the markers from a loss into a gain



Exact match

13.78%

vs 8.42% with tokens only,
+64% relative

Shuffle the <part_k> labels at test time

Jaccard	**0.459 → 0.209**
Exact match	**13.78% → 1.28%**

The LLM has bound each token to its own region.

RESULTS · PART-AWARE DESCRIPTION

Better part-aware descriptions on 3DCoMPaT-GrIn

Text track, multi-part split

Our model is trained without PartVerse-QA here, for a fair comparison

Method	BLEU-4	METEOR	SBERT	SimCSE	GPT-4o
ShapeLLM	8.01	30.05	**84.74**	85.38	–
PARIS3D	8.45	31.62	82.08	83.86	–
SegPoint	7.24	27.62	78.50	81.47	–
Kestrel	8.55	32.88	82.19	84.13	–
PointLLM (FT)	8.85	33.19	83.43	85.33	40.31
3D-PLOT-LLM (ours)	**9.32**	**33.93**	84.24	**86.23**	**43.34**

Bold marks the best value in each column.

+3.03

GPT-4o judge over PointLLM, with the same backbone and the same GrIn training data

Grounding without a decoder

Slots also give point masks. Zero-shot mIoU, single-part queries:

Our slot masks, no GrIn training



0.221

Best possible slot set on the 16 regions



0.448

RESULTS · WHOLE-OBJECT CAPTIONING

Whole-object captioning is preserved

Objaverse captioning

200 held-out objects

Method	Instruction data	SBERT	SimCSE	GPT-4o
PointLLM-PiSA	132K	48.21	48.38	–
ENEL	70K	47.70	48.22	37.02
PointLLM	70K	47.66	48.16	39.08
Ours, no PartVerse-QA	70K	47.94	48.44	40.56
3D-PLOT-LLM (ours)	147K	**48.31**	**48.67**	**40.93**

Bold marks the best value in each column. Our 147K is PointLLM's 70K plus 77K PartVerse-QA pairs.

Same 70K data as PointLLM

6 of 6

metrics in the paper are higher than PointLLM, without any PartVerse-QA

With PartVerse-QA

+1.85

GPT-4o judge over PointLLM, and +0.65 SBERT

TAKEAWAYS

Parts become tokens the LLM reads and writes

1. **Parts as vocabulary**

   Region tokens make parts addressable in both directions, with no learned detector, segmentation decoder, or box head, and under 1M new parameters.

2. **Supervision binds, MSR helps**

   PartVerse-QA ties the tokens to regions. MSR turns raw markers from a loss into a gain: C2S Jaccard 0.263 → 0.459.

3. **Coarse but sufficient**

   A fixed, training-free 16-region partition suffices for language-level part QA and preserves whole-object captioning. Point-wise masks still need a decoder.

Thank you

Paper
arxiv.org/abs/2606.19828

Code · Weights · Data
github.com/jintangxue/3D-PLOT-LLM