PLAN LAB
Perception and LANguage Lab

ELSA3D

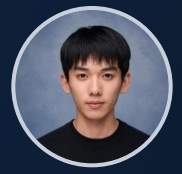
Elastic Semantic Anchoring for Unified 3D Understanding and Generation


Tianjiao Yu


Xinzhuo Li


Yifan Shen


Onkar Susladkar


Yuanzhe Liu


Xiaona Zhou


Ismini Lourentzou

University of Illinois Urbana-Champaign



NeurIPS 2026 / Atlanta Poster Session 5

UNIFIED 3D MODELS

One backbone for 3D generation and 3D understanding

Image → 3D
Reconstruct an object from a single image

Text → 3D
Generate an object from a description

3D → Text
Describe and reason about a shape in language

TODAY: ONE FLAT SEQUENCE, LEFT TO SELF-ATTENTION



▪ text tokens ▪ 3D tokens

Implicit interaction
Text–3D links must emerge from dense attention

Scales collapse
Coarse structure and fine detail share one stream

Fixed compute
Every layer runs at full width, for simple and complex inputs alike

TWO EXPLICIT TRACES

SEMANTIC TRACE

The description, split by abstraction level

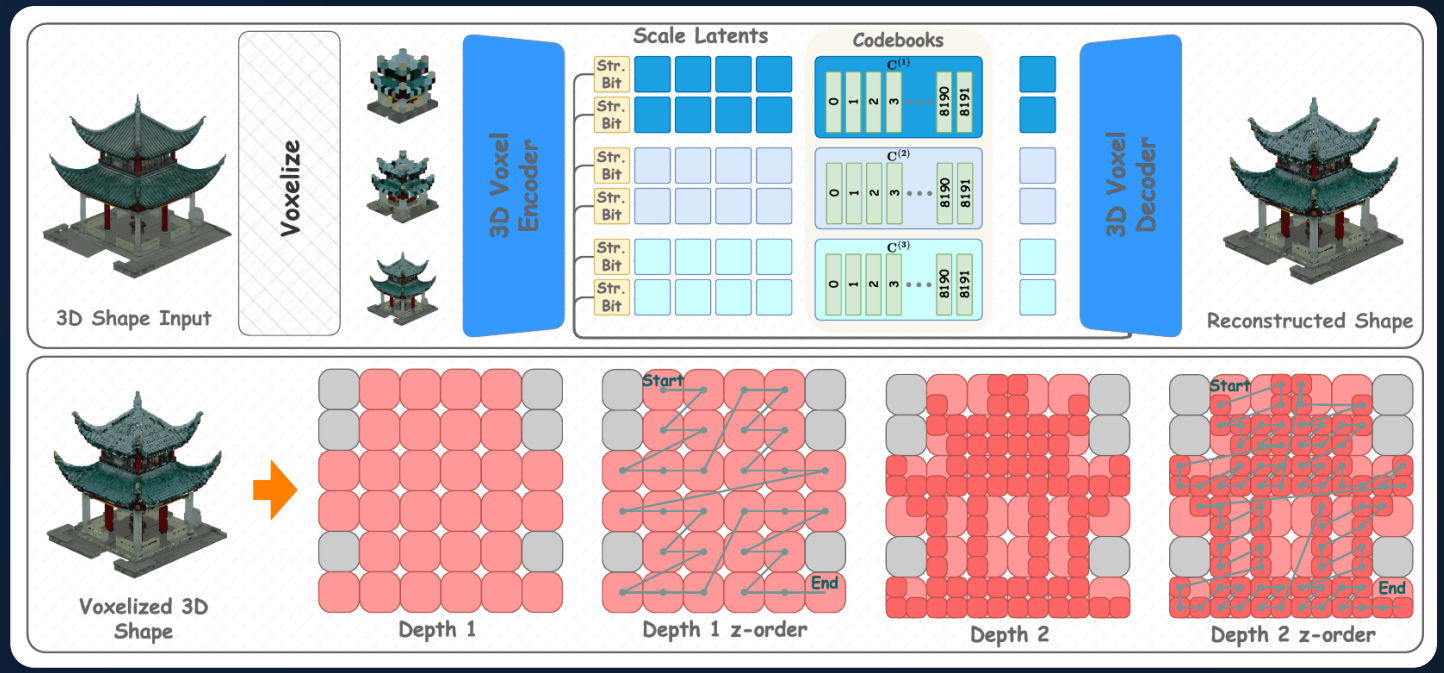
Global
category, silhouette, orientation

Structure
parts, proportions, mass

Appearance
material, color, texture

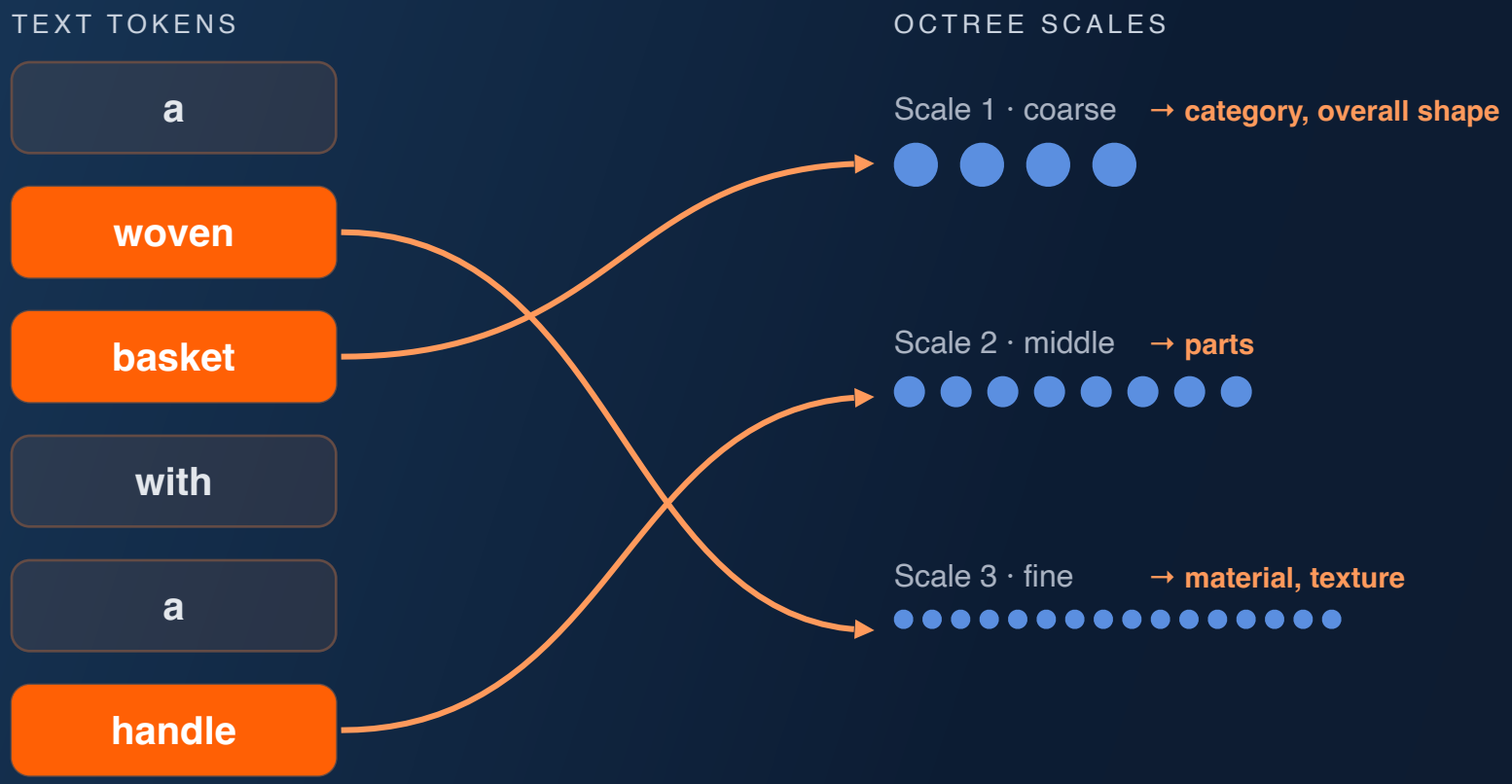
GEOMETRIC TRACE

Scale-aware octree VQ-VAE: a codebook per depth, a fixed scale tag on every token



ANCHOR TOKENS

A sparse, explicit bridge between language and geometry



Illustrative example. The router learns which words become anchors, and their scales, in every block.

Route & retrieve

Each anchor attends only to 3D tokens at its chosen scale

Fuse & write back

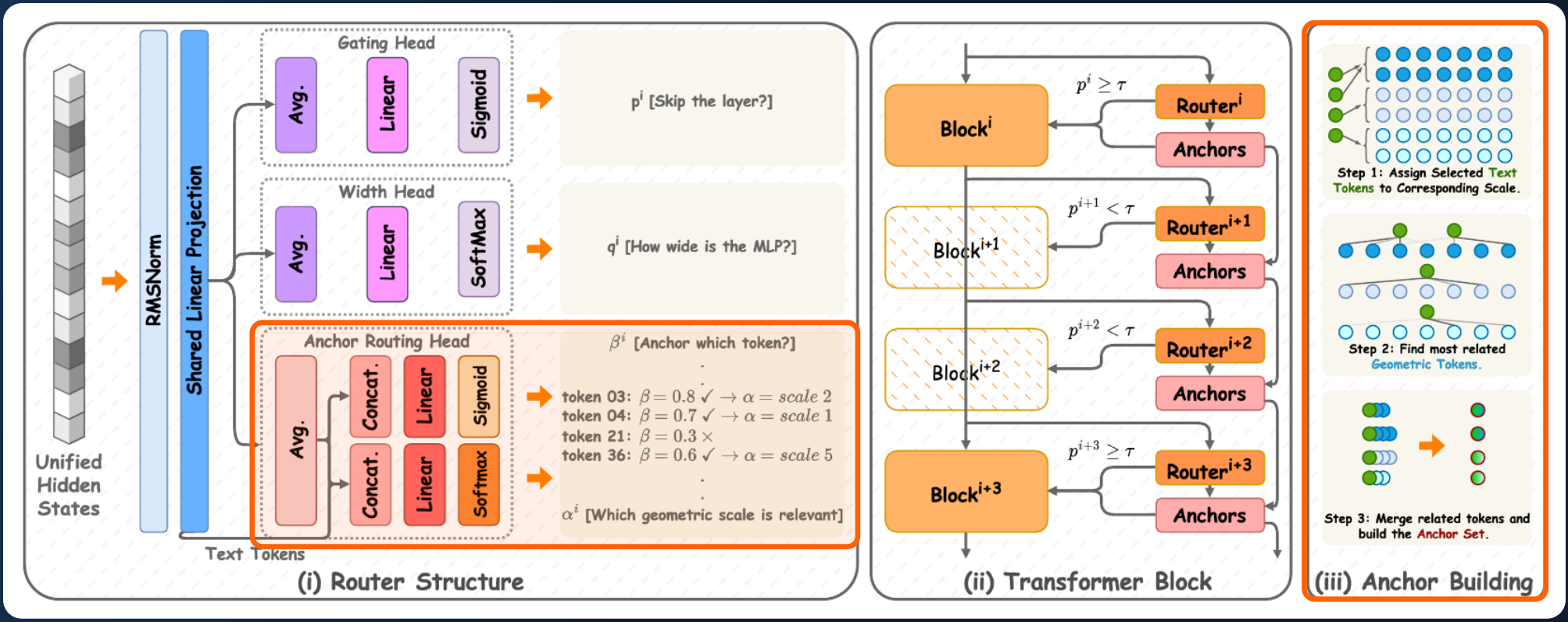
Gated cross-attention injects the anchor into the whole sequence



unified sequence, updated · anchors then discarded

ELASTIC ROUTER

A lightweight three-headed router on every transformer block



Gating head

Run or skip this block

Width head

MLP width: ¼, ½, ¾, or full

Anchor routing head

Which text tokens become anchors, and at which scale

TWO-STAGE TRAINING

STAGE 1

Scale-aware octree VQ-VAE

Learn discrete multi-scale geometry tokens from 3D data

STAGE 2

Unified autoregressive model

Freeze the tokenizer; train on interleaved text and 3D sequences, with auxiliary losses that shape the router

DATA

3D-Alpaca + Trellis-500K
3D assets with text and image pairs

UltraChat
Preserves general language ability

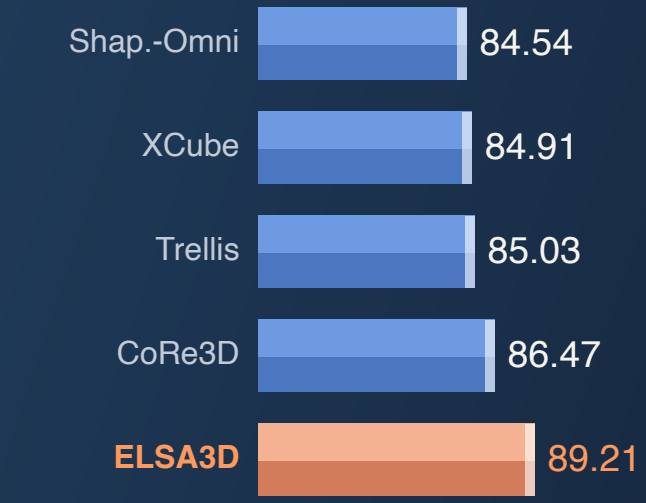
Toys4K + 200 in-the-wild images
Evaluation, with no training overlap

RESULTS ACROSS THREE TASKS

Compared with specialized 3D generators, vision-language models, and other unified 3D models

IMAGE-TO-3D

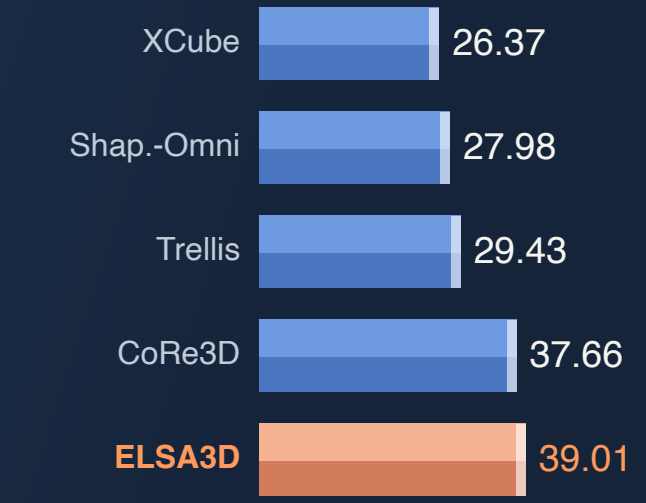
CLIP ↑ · axis from 70



Best on 5 of 7 metrics, and second-best on the other 2, among 12 baselines.

TEXT-TO-3D

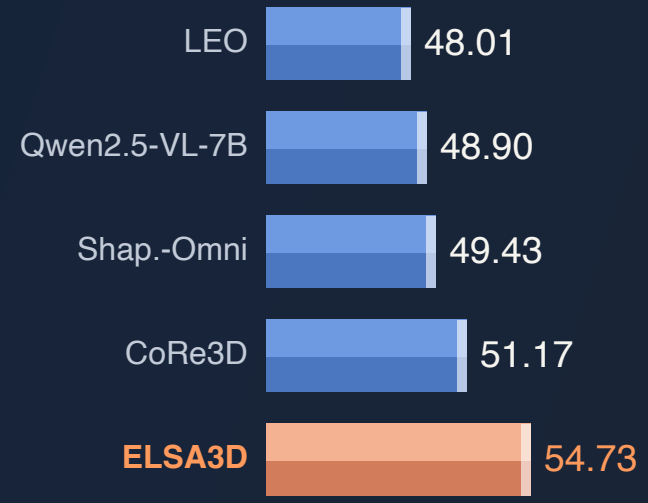
CLIP ↑



Best on all 4 metrics, among 9 baselines.

3D CAPTIONING

Sentence-BERT ↑ · axis from 40



Best on all 5 metrics, ahead of general and 3D-specialized models.

QUALITATIVE COMPARISON

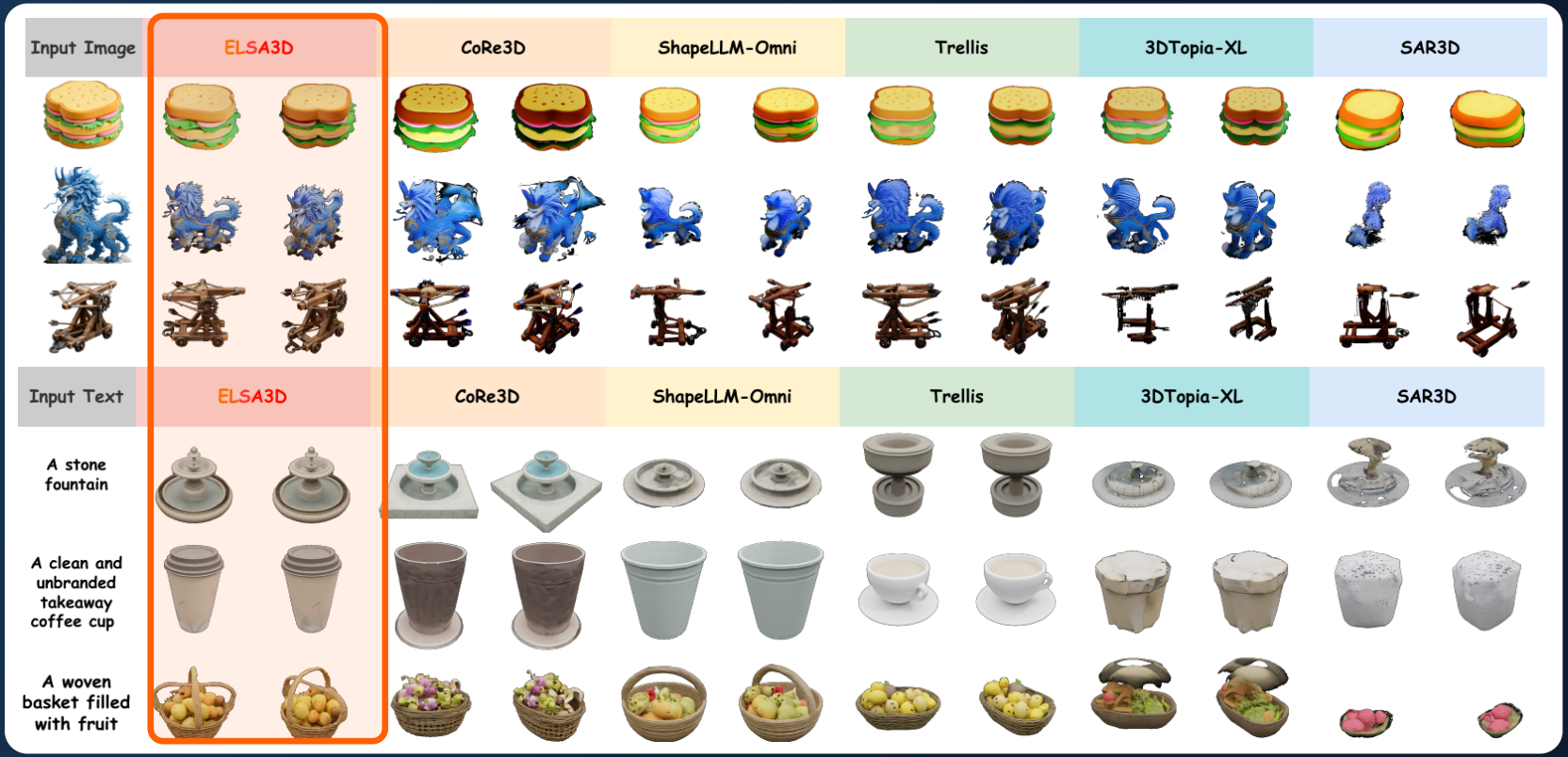


Image → 3D

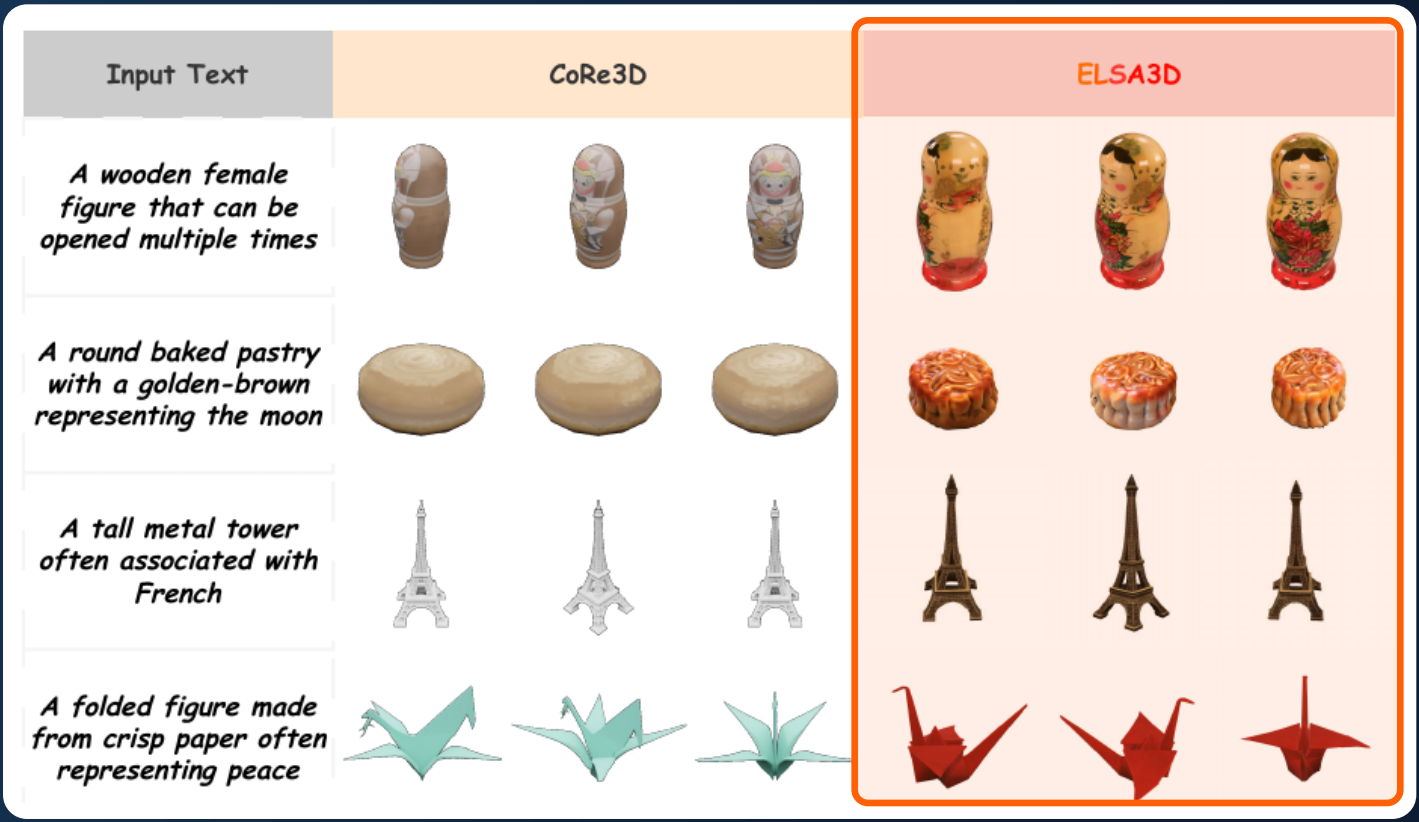
Keeps global shape and local appearance cues

Text → 3D

Follows fine constraints: a clean, unbranded cup; a basket that actually holds fruit

REASONING-BASED 3D GENERATION

Prompts describe the object indirectly, without naming it



"A wooden female figure that can be opened multiple times" → **nesting doll**

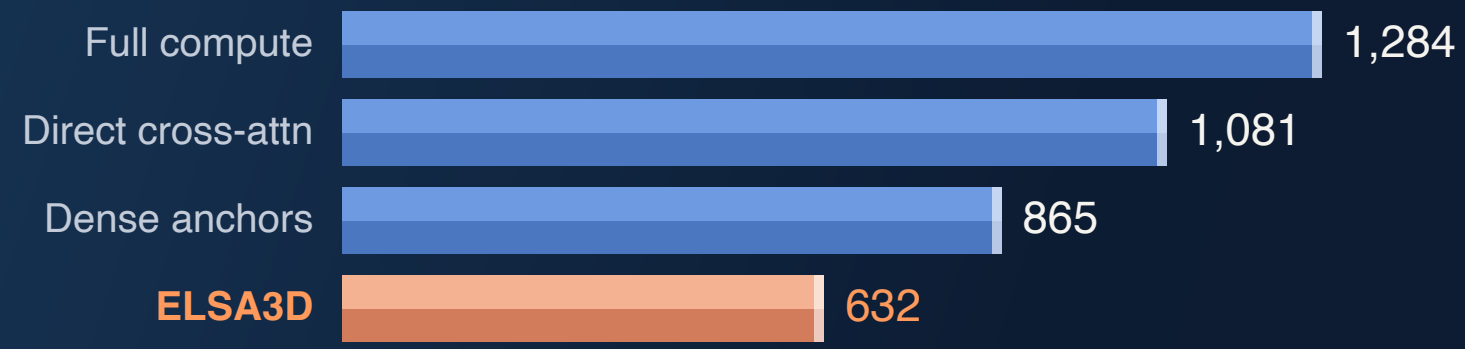
"A folded figure made from crisp paper, often representing peace" → **paper crane**

CoRe3D often returns only a coarse, generic shape.

ELASTIC COMPUTE

Same backbone, different interaction and compute designs

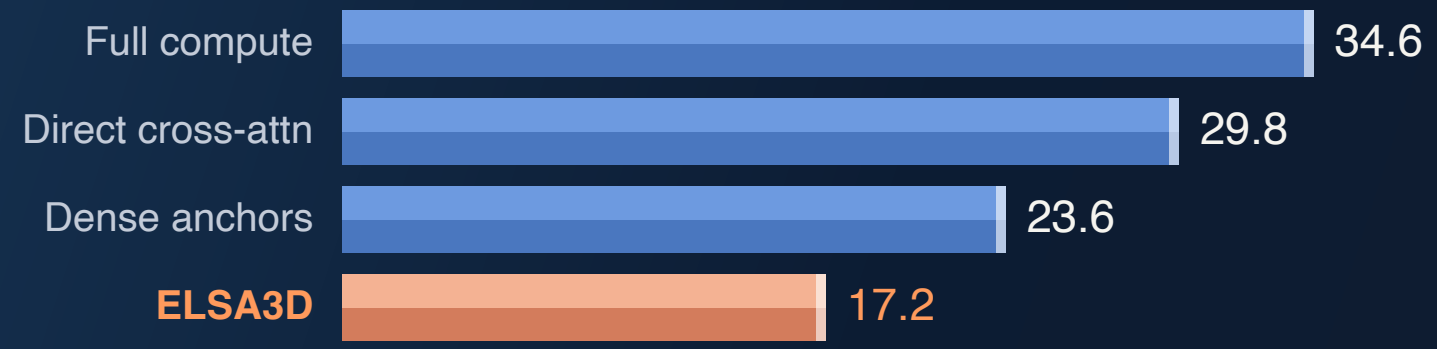
FLOPS (G) ↓



~½ FLOPs

of full compute, with near-identical quality: text-to-3D CLIP 39.01 vs 39.18

LATENCY (S) ↓

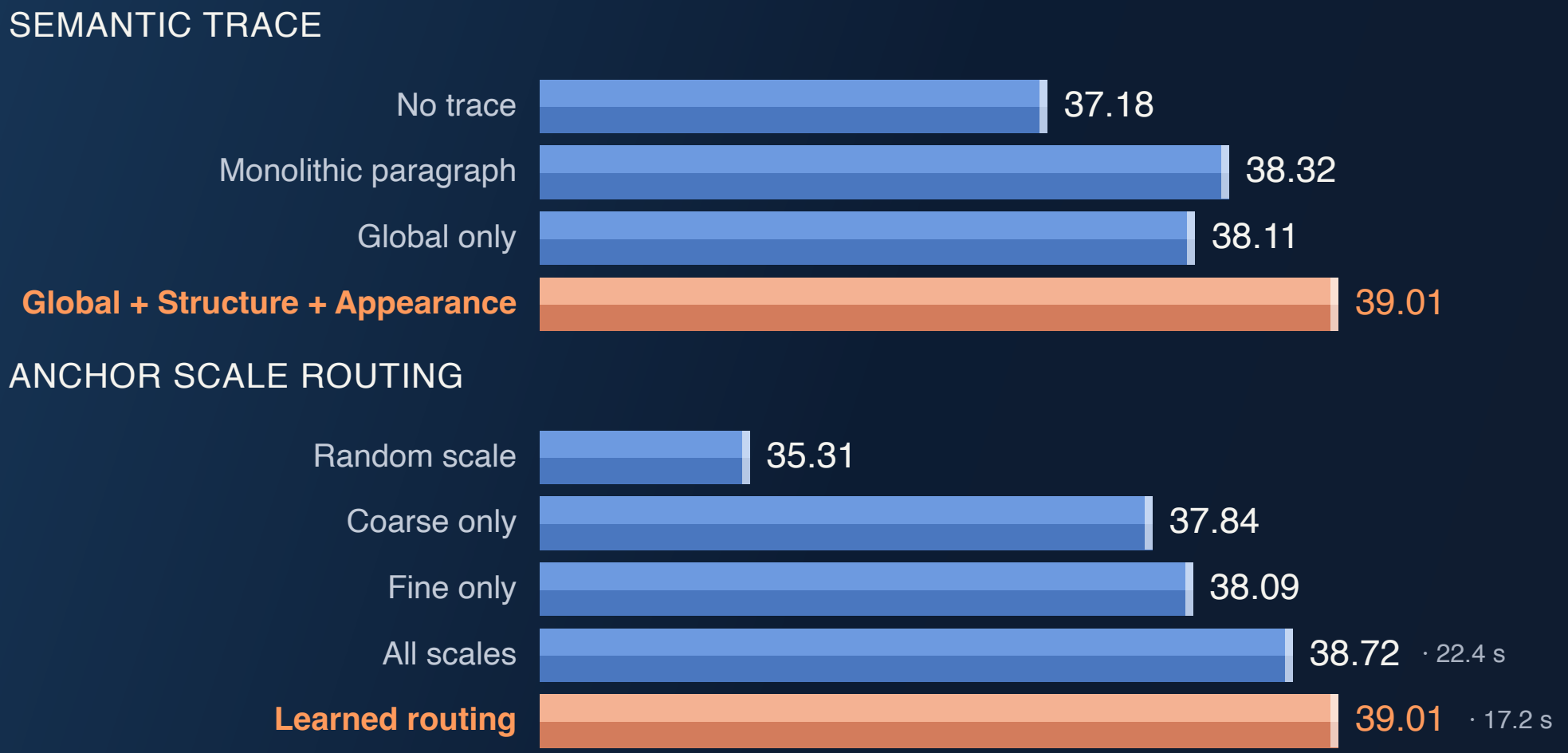


2× faster

than full compute; sparse anchors also beat dense fusion on quality

WHAT MATTERS

Text-to-3D CLIP ↑ (axis starts at 34)



SUMMARY

EXPLICIT TRACES

G · S · A + octree

Semantic cues and geometric scales, each made explicit

ANCHOR TOKENS

Sparse grounding

Selected words bind to evidence at one routed 3D scale

ELASTIC ROUTER

~½ FLOPs

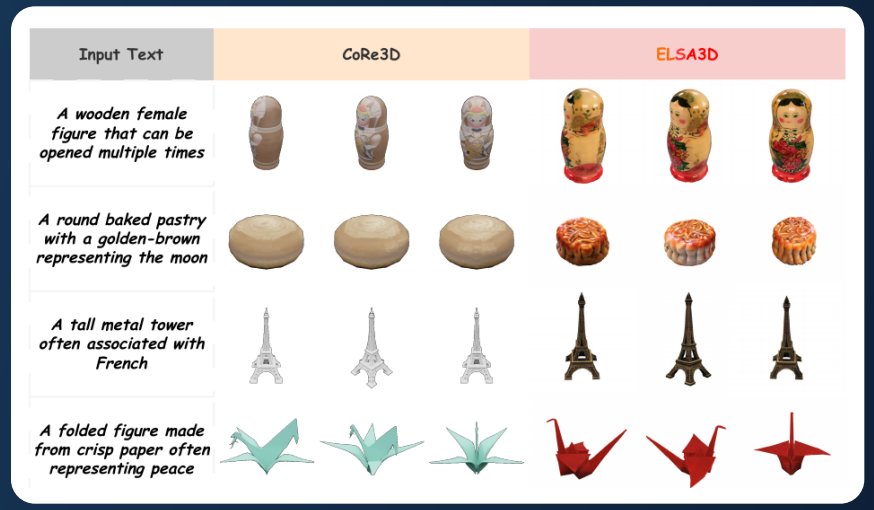
Strong results on image-to-3D, text-to-3D, and captioning

plan-lab.github.io/elsa3d

NEURIPS 2026 POSTER

Atlanta · Hall C1

PLAN LAB
Perception and LANguage Lab



THIS PAPER

ELSA3D

Elastic Semantic Anchoring for Unified 3D Understanding and Generation

FRIDAY, DECEMBER 11
10:00 AM ET

Hall C1 · Poster Session 5

PLAN LAB POSTERS AT NEURIPS 2026

WED, DEC 9 · 10:00 AM ET
Cognitive Supersensing
Toward Cognitive Supersensing in Multimodal Large Language Models
Poster Session 1

WED, DEC 9 · 10:00 AM ET
Tiny but Trusted
Efficient Vision-Language Reasoning for Time-Series Anomaly Detection
Poster Session 1

FRI, DEC 11 · 10:00 AM ET
ELSA3D
Elastic Semantic Anchoring for Unified 3D Understanding and Generation
Poster Session 5

FRI, DEC 11 · 4:30 PM ET
SILSA
Sliding-Window Slice Latents for Topology-Preserving High-Resolution 3D Generation
Poster Session 6

PLAN LAB
Perception and LANguage Lab

Thank you!



Tianjiao Yu

ty41@illinois.edu

plan-lab.github.io/elsa3d

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

ELSA3D · NeurIPS 2026

