NEURAL INFORMATION PROCESSING SYSTEMS

SCENEFORGE: Enhancing 3D-text alignment with Structured Scene Compositions

Cristian Sbrolli, Matteo Matteucci

Artificial Intelligence and Robotics Lab
Department of Electronics, Information and Bioengineering
Politecnico di Milano

POLITECNICO MILANO 1863

AIRLAB
ARTIFICIAL INTELLIGENCE AND ROBOTICS LAB

Introduction

The 3D Data Scarcity Problem

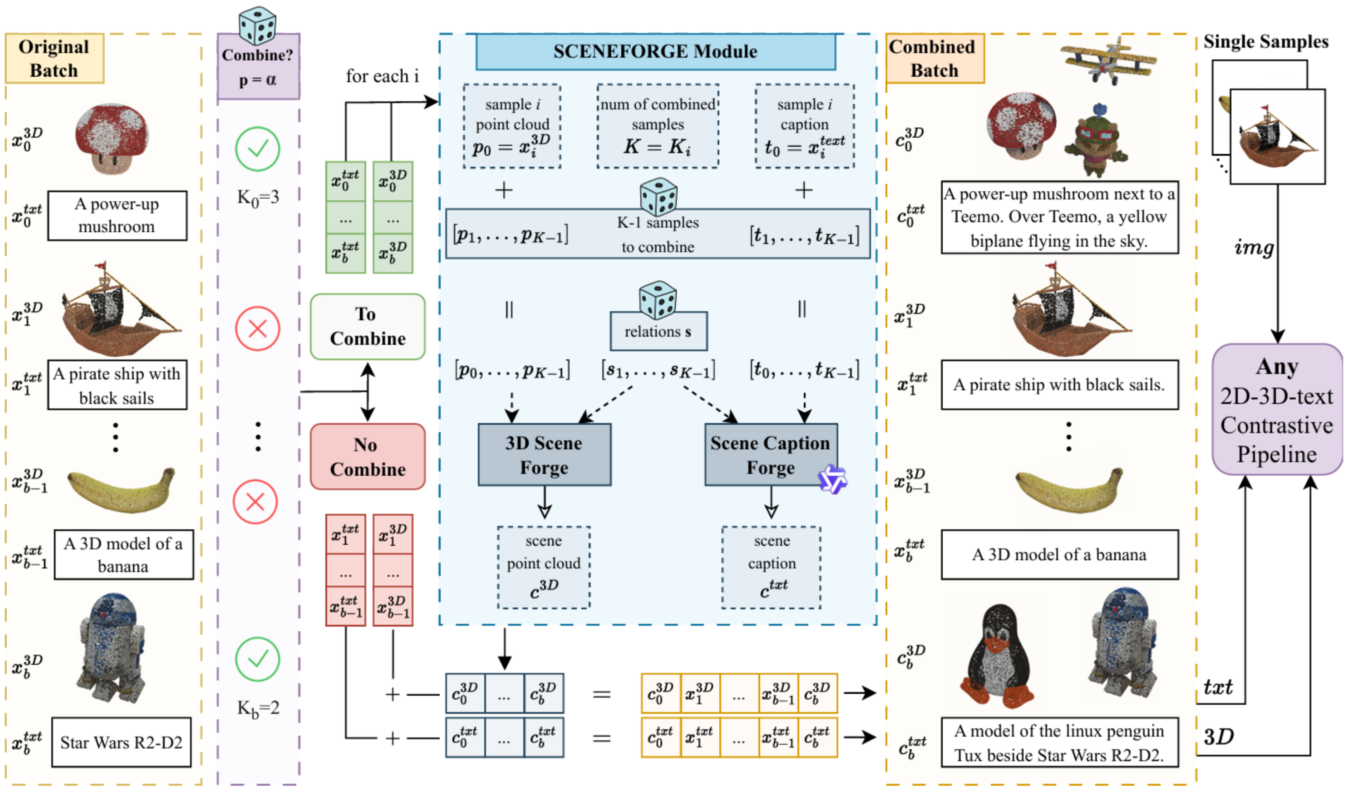
- Large-scale contrastive models work well by learning from **billions** of (image, text) pairs.
- This success is hard to replicate in 3D due to the **scarcity of large-scale 3D-text datasets**.
- Existing datasets are a great start but are still limited, focusing mostly on **single objects**.
- Real-world are **compositional**, defined by *multiple objects* and their *spatial relationships*.

Our Idea: "The whole is greater than the sum of its parts“

- **3D > 2D**: Unlike 2D images, 3D objects can be combined into complex scenes without visual artifacts.
- **Spatial Control**: We can explicitly control spatial relationships to create semantically harder scenes.
- Training on structured, multi-object scenes will teach the model richer, more robust representations.

Method

Large scale 3d datasets are only single object, how can we generalize to scenes?



- Combine single samples in scenes according to simple spatial relations: *“over”, “next to”*
- *Use an LLM to create a realistic scene caption for the composition.*
- *Combine each sample in the batch with a fixed probability.*
- *Combine up to N objects per sample.*

Method

How we combine point clouds and create scene captions:

Input: Samples p, Relations s, Target count P
Output: Composed 3D sample c^{3D}
$c^{3D}, p_{prev} \leftarrow \mathcal{A}^{3D}(p_0)$
for $i = 1$ *to* n **do**
  $p_i \leftarrow \mathcal{A}^{3D}(p_i)$
  $\Delta_{pos} \leftarrow \mathcal{P}(p_i, p_{prev}, s_i)$
  $p_i \leftarrow p_i + \Delta_{pos} + \delta + \epsilon$
  $c^{3D} \leftarrow \mathrm{cat}(c^{3D}, p_i)$
  $p_{prev} \leftarrow p_i$
end
$c^{3D} \leftarrow \mathcal{A}^{3D}(\mathrm{subsample}(c^{3D}, P))$
return c^{3D}

Algorithm 1: 3D Scene Forge algorithm.

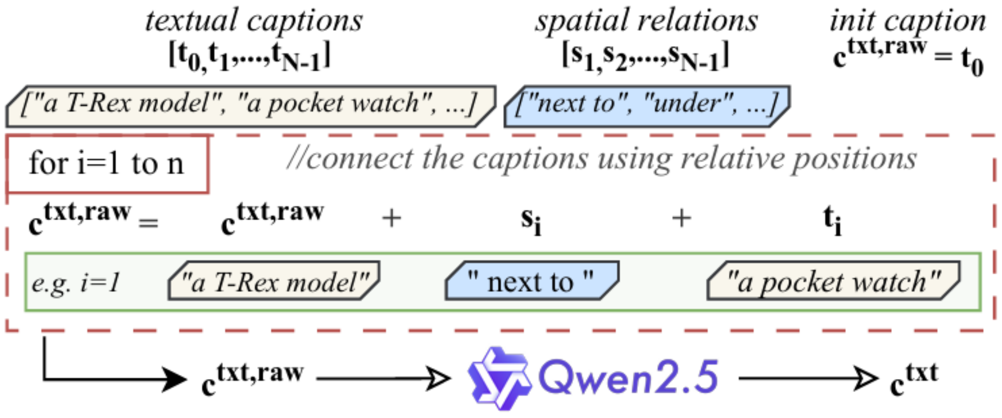


Figure 2: **Scene Caption Forge.** Starting from the initial caption (t_0), each caption (t_i) is connected using its relative position (s_i), creating a raw combined caption $c^{\text{txt,raw}}$. The raw caption $c^{\text{txt,raw}}$ is then refined to the final c^{txt} using Qwen2.5.

Method

Training loss with text-3D scene augmentation and 2D-3D singles:

Loss Partitioning. We consider contrastive models employing the InfoNCE loss proposed in CLIP [17]. For modalities $m, n \in \{txt, 2\mathrm{D}, 3\mathrm{D}\}$ and a sample subset $\mathcal{S}$, we define

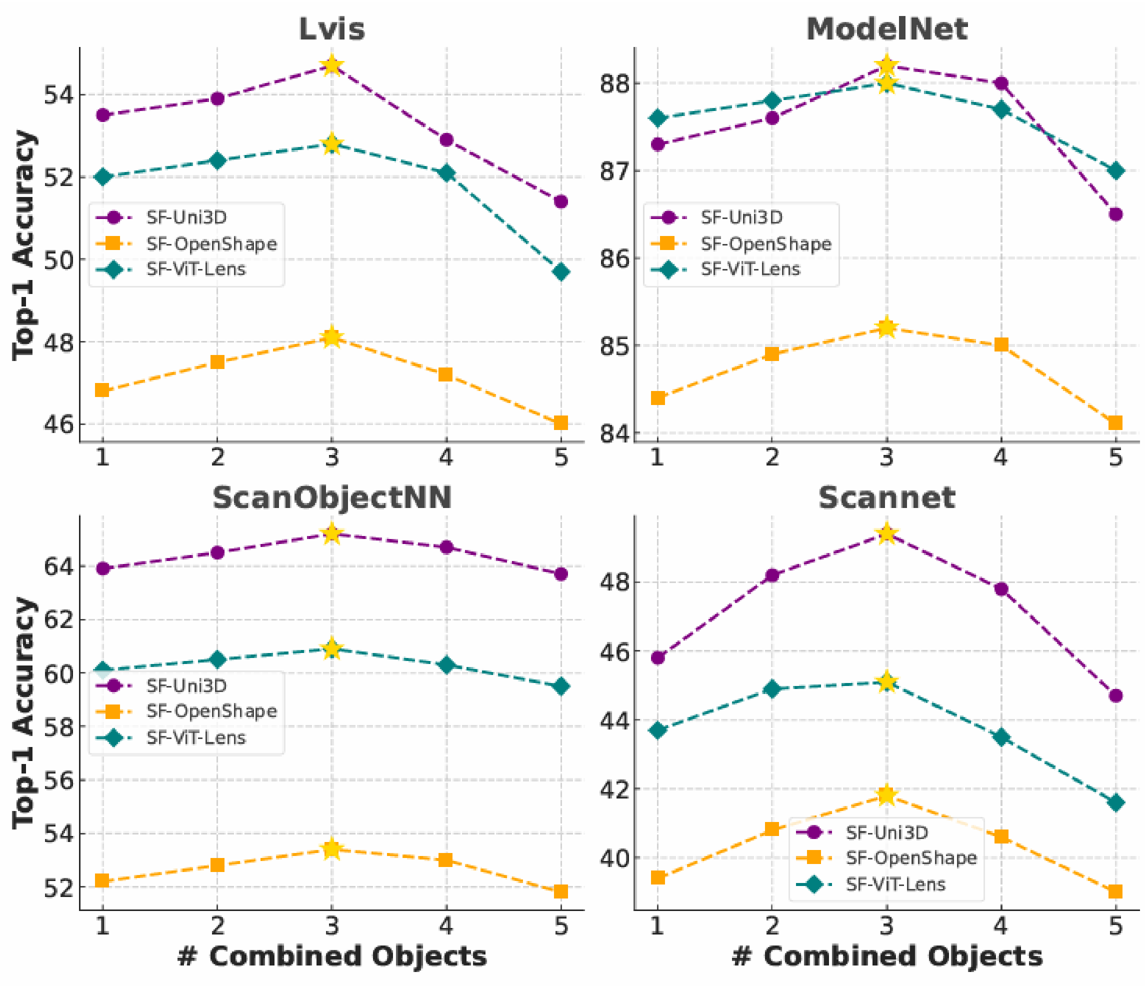
$$\mathcal{L}_{m \to n}(\mathcal{S}) = -\frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \log \frac{\exp\big(\langle e_i^m, e_i^n \rangle / \tau\big)}{\sum_{j \in \mathcal{S}} \exp\big(\langle e_i^m, e_j^n \rangle / \tau\big)}, \tag{1}$$

where e_i^m, e_i^n are ℓ_2-normalised embeddings and τ is a learnable temperature.

Let $\mathcal{S}_c$ and $\mathcal{S}_s$ denote the composed and single-object samples in a batch, with $N = |\mathcal{S}_c| + |\mathcal{S}_s|$. Because each sample is composed with probability α, $\mathbb{E}[|\mathcal{S}_s|] = (1 - \alpha)N$. We scale the image–3D block so that, *per batch*, it contributes the same total gradient budget as the text–3D block:

$$\mathcal{L} = \underbrace{\tfrac{1}{2}\big[\mathcal{L}_{3\mathrm{D} \to txt}(\mathcal{S}_c \cup \mathcal{S}_s) + \mathcal{L}_{txt \to 3\mathrm{D}}(\mathcal{S}_c \cup \mathcal{S}_s)\big]}_{\text{text–3D (all } N \text{ samples)}} + \frac{N}{|\mathcal{S}_s|} \underbrace{\tfrac{1}{2}\big[\mathcal{L}_{3\mathrm{D} \to 2\mathrm{D}}(\mathcal{S}_s) + \mathcal{L}_{2\mathrm{D} \to 3\mathrm{D}}(\mathcal{S}_s)\big]}_{\text{2D–3D (singles only)}}. \tag{2}$$

Results



What is the optimal value of the maximum #objects in a scene?

Evaluation is performed on established single-object and scene benchmarks, not on our scene compositions.

Results

Detailed results for best value N=3

Classification

Model	LVIS T1	LVIS T5	ModelNet T1	ModelNet T5	ScanObjNN T1	ScanObjNN T5	Scannet T1	Avg Δ
ULIP 2	46.3	75.0	84.0	97.2	45.6	82.9	38.1	–
TAMM	42.0	71.7	86.3	98.1	56.7	86.1	42.4	–
MixCon3D	47.5	76.2	87.3	98.1	57.7	89.8	43.0	–
OmniBind-L	–	–	–	–	–	–	–	–
OmniBind-F	–	–	–	–	–	–	–	–
OpenShape	39.1	68.9	85.3	97.4	47.2	84.7	40.3	+1.50
SF-OpenShape	41.7	71.5	86.7	98.1	48.0	85.9	41.5	
ViT-Lens	50.1	78.1	86.8	97.8	59.8	87.7	43.8	+0.78
SF-ViT-Lens	**50.9**	**78.4**	87.3	98.0	60.9	89.1	44.5	
Uni3D	47.2	76.1	86.8	98.4	66.5	90.1	43.9	+1.73
SF-Uni3D	48.9	**78.4**	**87.5**	**99.0**	**67.3**	**91.5**	**47.6**	

(a) Trained on ensemble (no LVIS).

Model	LVIS T1	LVIS T5	ModelNet T1	ModelNet T5	ScanObjNN T1	ScanObjNN T5	Scannet T1	Avg Δ
ULIP 2	50.6	79.1	84.7	97.1	51.5	89.3	38.9	–
TAMM	50.7	80.6	85.0	98.1	55.7	88.9	41.8	–
MixCon3D	52.5	81.2	86.8	98.3	58.6	89.2	44.1	–
OmniBind-L	54.0	82.9	86.6	99.0	64.7	94.2	46.3	–
OmniBind-F	53.6	81.8	87.1	99.0	64.7	**94.4**	46.1	–
OpenShape	46.8	77.0	84.4	98.0	52.2	88.7	39.4	+1.43
SF-OpenShape	48.1	78.4	85.2	98.3	53.4	89.5	41.8	
ViT-Lens	52.0	79.9	87.6	98.4	60.1	90.3	43.7	+0.85
SF-ViT-Lens	52.8	80.7	88.0	89.9	60.9	91.2	45.1	
Uni3D	53.5	82.0	87.3	**99.2**	63.9	91.7	45.8	+1.75
SF-Uni3D	**54.7**	**84.8**	**88.2**	**99.2**	**65.2**	93.4	**49.4**	

(b) Trained on ensemble (with LVIS).

Few-shot segmentation & 3D VQA

Method	1-shot mIoU	1-shot Δ	2-shot mIoU	2-shot Δ
OmniBind-L	77.2	–	79.9	–
OmniBind-F	77.8		80.3	
OpenShape	74.0	+2.2	76.5	+2.6
SF-OpenShape	76.2		79.1	
ViT-Lens	75.5	+1.5	77.9	+2.2
SF-ViT-Lens	77.0		80.1	
Uni3D	75.9	+2.6	78.2	+3.0
SF-Uni3D	**78.5**		**81.2**	

Table 2: One-shot and two-shot part segmentation on ShapeNetPart.

Model	B-4	ΔB-4	CIDEr	ΔCIDEr	EM	ΔEM
OmniBind-L + BLIP2-FlanT5	8.5	–	62.9	–	17.1	–
OmniBind-F + BLIP2-FlanT5	8.3		62.1		17.6	
OpenShape + BLIP2-FlanT5	6.3	+1.8	54.8	+6.7	14.1	+2.8
SF-OpenShape + BLIP2-FlanT5	8.1		61.5		16.9	
ViT-Lens + BLIP2-FlanT5	7.2	+1.3	57.5	+5.9	15.7	+2.1
SF-ViT-Lens + BLIP2-FlanT5	8.5		63.4		17.8	
Uni3D + BLIP2-FlanT5	7.5	+2.9	58.3	+8.4	16.4	+4.1
SF-Uni3D + BLIP2-FlanT5	**10.4**		**66.7**		**20.5**	

Table 3: Performance on the ScanQA dataset using BLEU-4, CIDEr, and Exact Match.

Results

Our models also behave better when used as initialization for fine tuning on classification benchmarks with peft methods.

Table 4: Supervised fine-tuning accuracy (%).

Model	Method	Trainable Params	ModelNet40	ScanObjectNN	ScanNet Inst.
Uni3D	Full Fine-Tuning	1016.5M (100%)	94.28	97.12	82.72
	Adapter	7.6M (0.74%)	94.35	96.80	81.42
	DAPT	7.3M (0.72%)	94.33	96.78	82.65
	PointGST	4.1M (0.40%)	94.83	97.68	83.04
SF-Uni3D	Full Fine-Tuning	1016.5M (100%)	94.42	97.58	83.58
	Adapter	7.6M (0.74%)	94.46	97.09	82.56
	DAPT	7.3M (0.72%)	94.49	97.15	83.46
	PointGST	4.1M (0.40%)	**94.95**	**98.09**	**84.29**

Results

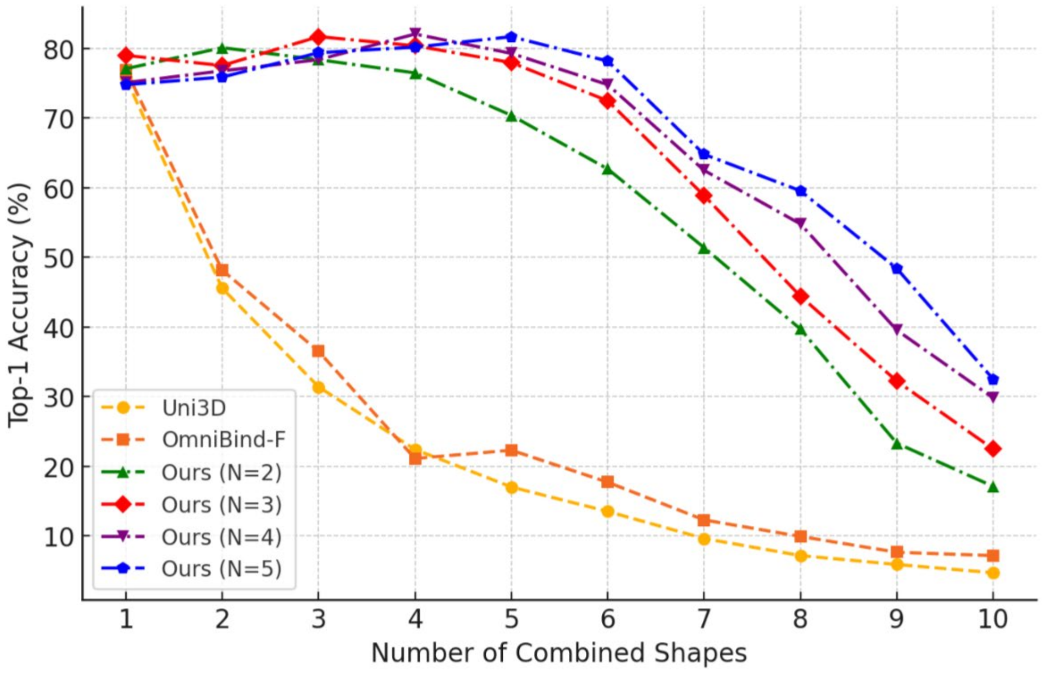


Figure 4: Top-1 averaged retrieval accuracy on the N-LVIS datasets as N increases.

How well does each model generalize to more complex scenes?

Evaluation is performed on the N-LVIS dataset, where each sample is a combination of N different ones.

Results

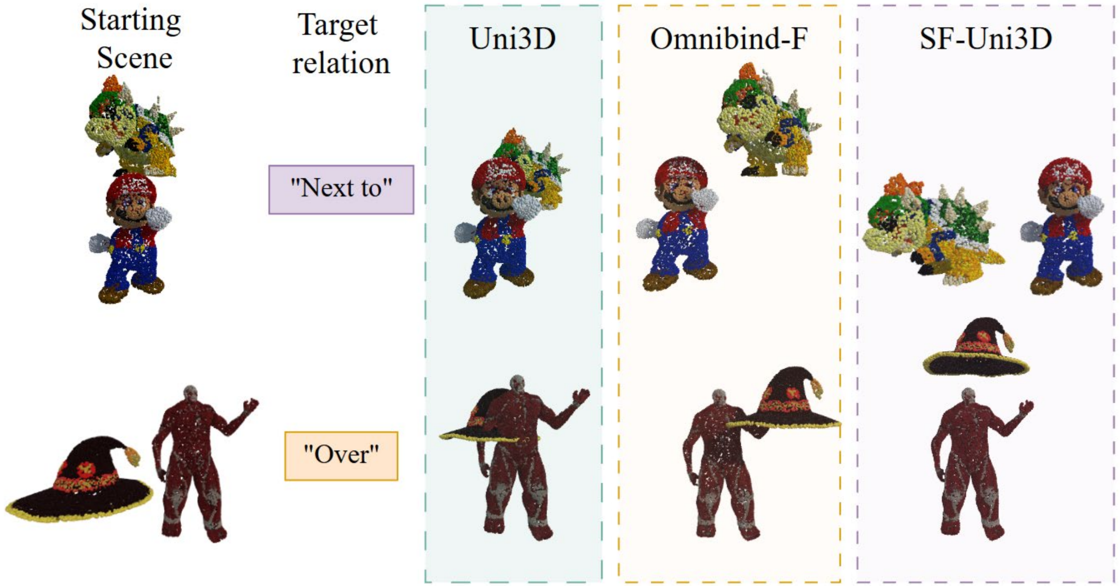


Figure 5: Object repositioning example.

Qualitatively verifying that the model has learnt spatial semantics

A tridimensional offset vector is optimized for the second object with frozen encoders, maximing cosine similarity (regularized).

$$\mathcal{L} = -\cos(\phi(x^{3D} + \Delta), \psi(x^{txt})) + \lambda||\Delta||^2$$

Ablations

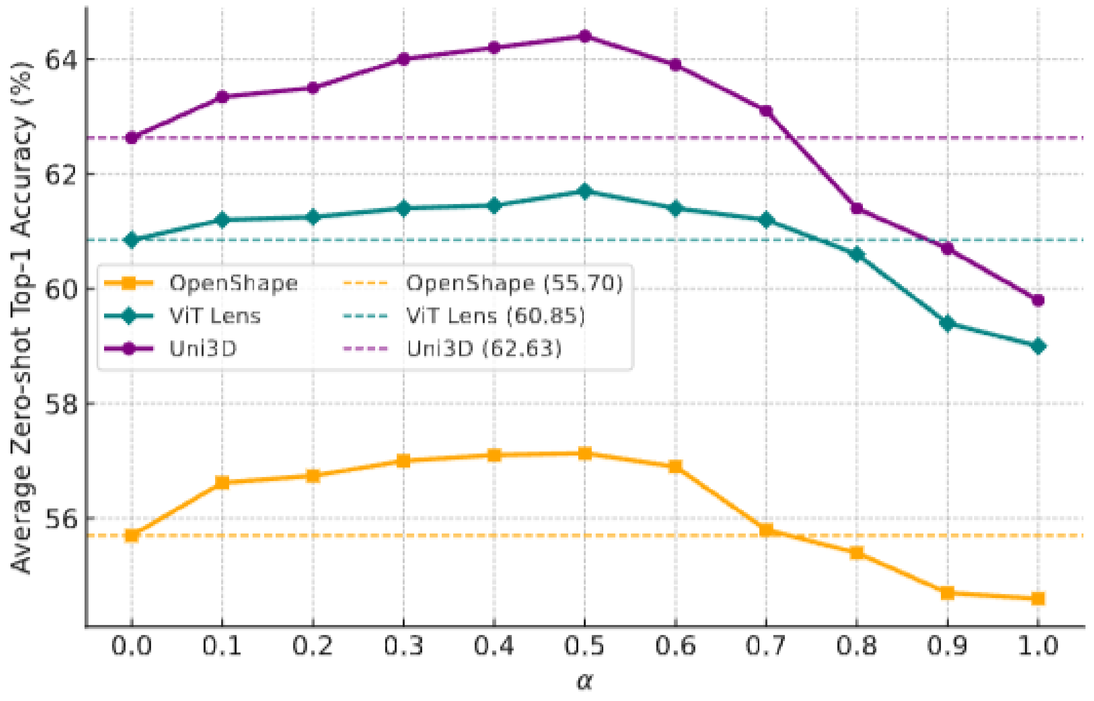


Figure 6: Effect of varying α on average zero-shot top-1 accuracy.

What is the best value for alpha?

Zero-Shot accuracy is averaged over the considered benchmarks.

Ablations

Do other composition functions work?

PointCutMix-K, the only one preserving key features of mixed shapes, is the only one able to bring improvements, confirming the importance of well-structured compositions.

Composition Method	Lvis Top-1	ModelNet Top-1	ScanObjNN Top-1	Scannet Top-1
None (Uni3D)	*53.5*	*87.3*	63.9	45.8
PointCutMix-K	*53.5*	87.1	*64.1*	*47.5*
PointCutMix-R	44.7	83.0	45.1	34.8
PointMixup	39.2	78.7	41.4	30.2
SF-Uni3D (N=2)	**53.9**	**87.6**	**64.5**	**48.2**

Table 5: Different 3D composition methods on zero-shot cls.

Ablations

Table 6: Generalization to unseen spatial relations on ScanQA for all backbones.

Relation Type	Metric	OpenShape			ViT-Lens			Uni3D		
		Baseline	SF	Δ	Baseline	SF	Δ	Baseline	SF	Δ
Attached To (21)	CIDEr	54.5	61.5	+7.0	57.1	63.3	+6.2	57.9	**66.6**	+8.7
	EM	14.1	17.1	+3.0	15.6	17.9	+2.3	16.5	**20.8**	+4.3
Sitting On (59)	CIDEr	56.8	63.4	+6.6	59.0	65.1	+6.1	61.0	**70.1**	+9.1
	EM	15.2	17.7	+2.5	16.6	18.4	+1.8	17.5	**22.6**	+5.1
Between (112)	CIDEr	54.1	61.2	+7.1	56.8	62.9	+6.1	57.2	**66.5**	+9.3
	EM	14.0	17.0	+3.0	15.5	17.9	+2.4	15.8	**20.5**	+4.7
Closest To (112)	CIDEr	55.0	61.8	+6.8	57.5	63.5	+6.0	58.5	**67.0**	+8.5
	EM	14.3	17.2	+2.9	15.8	18.0	+2.2	16.2	**20.4**	+4.2
In Front Of (246)	CIDEr	56.1	62.5	+6.4	58.2	64.0	+5.8	60.3	**68.3**	+8.0
	EM	14.9	17.6	+2.7	16.3	18.2	+1.9	17.1	**21.8**	+4.7

Are simple relations enough?

We isolate questions from ScanQA involving relations unseen during training, and we show that our models correctly generalize and improve VQA also in these cases.

Thank You!

corresponding author:
cristian.sbrolli@polimi.it