

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

Cristian Sbrolli

Department of
Electronics, Information and Bioengineering
Politecnico di Milano
Via Ponzio 34/5, 20133 Milan, Italy
cristian.sbrolli@polimi.it

Matteo Matteucci

Department of
Electronics, Information and Bioengineering
Politecnico di Milano
Via Ponzio 34/5, 20133 Milan, Italy
matteo.matteucci@polimi.it

N	LVIS	M40	ScanObjNN	ScanNet	# Comb.	LVIS	M40	ScanObjNN	ScanNet	# Comb.	LVIS	M40	ScanObjNN	ScanNet
1	53.5	87.3	63.9	45.8	1	46.8	84.4	52.2	39.4	1	52.0	87.6	60.1	43.7
2	53.9	87.6	64.5	48.2	2	47.5	84.9	52.8	40.8	2	52.4	87.8	60.5	44.9
3	54.7	88.2	65.2	49.4	3	48.1	85.2	53.4	41.8	3	52.8	88.0	60.9	45.1
4	52.9	88.0	64.7	47.8	4	47.2	85.0	53.0	40.6	4	52.1	87.7	60.3	43.5
5	51.4	86.5	63.7	44.7	5	46.0	84.1	51.8	39.0	5	49.7	87.0	59.5	41.6

(a) SF-Uni3D
(b) SF-OpenShape
(c) SF-ViT-Lens

Table 1: Top-1 Accuracy varying N , extended results.

1 Optimal value of N : Extended Results

The choice of N , the number of combined objects, significantly impacts performance across all evaluated benchmarks, as shown in table 1. The results indicate that increasing N initially improves accuracy but may lead to diminishing returns or slight degradation beyond a certain point. For all datasets, the highest Top-1 accuracy is consistently achieved at $N = 3$, suggesting that moderate object combination enhances recognition without introducing excessive ambiguity or complexity. In particular, ModelNet reaches its peak Top-1 accuracy of 88.2%, while Lvis and ScanObjNN achieve their best performances at 54.7% and 65.2%, respectively. Scannet exhibits a similar trend, peaking at 49.4%. The Top-5 accuracy follows a comparable pattern, where $N = 3$ yields the best overall results for Lvis (84.8%) and ModelNet (99.2%). However, for ScanObjNN, the highest Top-5 accuracy is observed at $N = 2$ (93.8%), and ModelNet achieves its best Top-5 score at $N = 4$ (99.3%), indicating slight variations in optimal N across datasets. Notably, performance starts to decline beyond $N = 3$, particularly for Lvis and ScanObjNN, where Top-1 scores decrease at $N = 4$ and $N = 5$. This suggests that excessive object combination may introduce challenges related to occlusion, feature confusion, or increased intra-class variance, ultimately hampering recognition performance. Overall, these results highlight $N = 3$ as the most effective choice, balancing robustness and discriminability across multiple datasets. While dataset-specific variations exist, selecting an optimal N is crucial for maximizing recognition accuracy in multimodal learning settings.

1.1 Caption Refinement Prompt

To ensure that the textual descriptions of our generated composite scenes are natural, well-structured, and grammatically correct, we employ a large language model (*Qwen 2.5 7B-instruct* [4]) to refine the raw captions. The raw captions are constructed by sequentially concatenating object descriptions with their respective spatial relations (e.g., “next to,” “over,” “under”), resulting in text that often contains structural inconsistencies, punctuation artifacts, and unnatural phrasing.

Methods	LVIS	ModelNet	ScanObjNN	Scannet
Uni3D	53.5	87.3	63.9	45.8
SF-Uni3D/NO LLM	53.8	87.8	64.7	48.3
SF-Uni3D/Llama 3.1	54.6	88.2	65.3	49.2
SF-Uni3D/Qwen 2.5	54.7	88.2	65.2	49.4

Table 2: We evaluate the impact of the LLM model used for caption refinement.

To address these issues, we utilize the following prompt to guide the LLM in transforming the raw caption into a coherent and fluent description:

You are an advanced language model specializing in natural language refinement. Your task is to transform a raw combined caption into a fluent and well-structured description. The raw caption consists of multiple object descriptions connected by spatial relations such as “next to,” “over,” or “under.” Your goal is to ensure proper grammar, capitalization, and punctuation while making the text more readable and natural. If the sentence is too long or unnatural, split it into multiple fluent sentences while preserving the original meaning and spatial relationships. The final output should feel like a human-written description, maintaining clarity and coherence.

Here is the raw caption:

Raw Caption: {insert raw caption here}

Refined Caption:

This prompt explicitly instructs the LLM to focus on linguistic refinement without altering the underlying semantics of the scene description. By enforcing grammatical correctness and natural phrasing, it ensures that the final captions effectively communicate the spatial composition of the 3D scenes while maintaining readability and coherence.

2 Impact of Caption Refinement

The effectiveness of our Multimodal Scene Crafter (MSC) relies not only on the composition of 3D objects but also on the quality of the corresponding textual descriptions. To assess the impact of refining the raw captions generated by MSC, we compare three training settings on the best variant, SF-Uni3D: (i) the baseline Uni3D [5], where individual objects ($N = 1$) are processed without composition, (ii) the use of raw combined captions, and (iii) the use of LLM-refined captions.

As shown in table 2, incorporating multiple objects into a single scene already improves performance over the Uni3D baseline, even when using unprocessed raw captions. However, the raw captions often contain structural inconsistencies and artifacts that limit their effectiveness as training signals. By refining the captions with a large language model (LLM), we significantly enhance the clarity and coherence of scene descriptions, leading to notable improvements in zero-shot classification accuracy.

To further investigate the impact of the LLM choice, we evaluated both Llama 3.1 8B [1] and Qwen 2.5-Instruct [4] for caption refinement. As seen in table 2, both models yielded similar performance gains, with Qwen 2.5-Instruct slightly outperforming Llama across most datasets. Given this marginal improvement, we opted to use Qwen 2.5-Instruct for all subsequent experiments. These results underscore the importance of linguistic quality in vision-language contrastive learning, as refined captions improve the alignment between 3D scene representations and textual descriptions, enabling the model to better capture spatial relationships and object interactions.

3 Theoretical Scene Cardinality

We analyze the theoretical number of unique scenes that can be constructed given:

- A dataset of D distinct objects.
- A scene containing up to N objects.

Method	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr	EM
ULIPPointBERT + FlanT5	29.2	17.9	10.3	6.1	11.6	28.1	50.9	14.5
OmniBind-Large + BLIP2-FlanT5	32.7	21.7	14.2	8.5	13.1	32.4	62.9	17.1
OmniBind-Full + BLIP2-FlanT5	33.0	22.3	14.5	8.3	13.6	33.6	62.1	17.6
OpenShape + BLIP2-FlanT5	28.5	18.2	11.2	6.3	11.1	29.8	54.8	14.1
SF-OpenShape + BLIP2-FlanT5	31.0	20.5	13.0	8.1	12.8	32.9	61.5	16.9
ViT-Lens + BLIP2-FlanT5	30.0	19.2	11.8	7.2	12.1	30.7	57.5	15.7
SF-ViT-Lens + BLIP2-FlanT5	32.5	21.5	14.0	8.5	13.3	33.8	63.4	17.8
Uni3D + BLIP2-FlanT5	31.2	20.9	13.1	7.5	12.4	30.1	58.3	16.4
SF-Uni3D + BLIP2-FlanT5	35.9	23.2	15.1	10.4	14.6	35.4	66.7	20.5

Table 3: Extended VQA results (BLEU-1–4, METEOR, ROUGE-L, CIDEr, Exact Match).

- Objects cannot repeat within a scene.
- Object order matters.
- Each object is related only to its preceding object, leading to $k - 1$ relations in a scene of size k .

For a scene of size k , the number of ways to select and arrange k objects from D without replacement and order-sensitive (as then spatial relations are affected by order) is:

$$P(D, k) = \frac{D!}{(D - k)!}. \quad (1)$$

Each of the $k - 1$ relations can take one of R possible values, leading to R^{k-1} possible configurations per group of k objects. Summing over all possible scene sizes from 1 to N , the total number of unique compositional scenes is:

$$S(N, D, R) = \sum_{k=1}^N P(D, k) R^{k-1}. \quad (2)$$

We evaluate the number of possible scenes for our best case, $N=3$, $S(3, 876000, 3)$. Substituting the given values:

$$S(3, 876000, 3) \approx 6E. \quad (3)$$

where E denotes **exa** (10^{18}). This result underscores the combinatorial explosion in scene generation, even with a small number of objects per scene. Given the factorial and exponential growth of $S(N, D, R)$, exhaustively enumerating all possible scenes is infeasible, which may result in the omission of meaningful configurations during training while including unrealistic ones. A promising future direction is then to enhance scene generation by incorporating semantic-aware relation and object selection. Instead of random sampling, leveraging structured priors to enforce plausible spatial and functional interactions could lead to more coherent and realistic scene configurations, better reflecting real-world distributions.

4 Additional Results on ScanQA

In the main paper, we focused on BLEU-4, CIDEr, and Exact Match for 3D question answering on ScanQA, while here we provide extended metrics (BLEU-1, BLEU-2, BLEU-3, METEOR, ROUGE-L) to offer a more comprehensive view of each method’s linguistic performance. Following the protocol of prior contrastive 3D encoders [2], we use Uni3D-giga, OmniBind-Large, OmniBind-Full, and our approach as 3D feature extractors connected to BLIP2-FlanT5 [3]. Table 3 summarizes these extended results, where our method again achieves superior scores across all metrics. We attribute

Instances	AWQ	OpenShape	ViT-Lens	Uni3D
1	✗	5.00	1.10	0.30
2	✗	2.00	0.05	0.00
4	✗	0.50	0.00	0.00
1	✓	3.00	0.70	0.18
2	✓	1.00	0.00	0.00
4	✓	0.00	0.00	0.00

Table 4: Computational overhead of SceneForge with different backbones and quantization settings. Values represent multiplicative overhead relative to standard training time.

Model	Pix3D	Objaverse LVIS
OpenShape	0.512	65.706
SF-OpenShape	0.495	65.010
ViT-Lens	0.614	72.051
SF-ViT-Lens	0.622	71.203
Uni3D	0.641	74.840
SF-Uni3D	0.627	74.414

Table 5: Image–3D retrieval performance measured by average top-1 accuracy (higher is better) on Pix3D and Objaverse LVIS.

these improvements primarily to our combined-sample pretraining, which we believe promotes deeper cross-instance reasoning and sharper recognition of spatial relationships. As noted in the main text, while other models handle color- or attribute-based questions adequately, ours shows a clearer advantage when queries hinge on object-object interactions (e.g., “What is over the brown chair?”). This suggests that our approach not only refines local feature encoding but also better infers complex 3D layouts, a core strength for scene-level comprehension.

5 Computational Overhead of the SceneForge Pipeline

We analyze the computational overhead introduced by the SceneForge pipeline, particularly due to the LLM-based caption refinement module. To quantify this overhead, we measure the ratio between the batch processing time (training a single batch) and the additional time needed for SceneForge’s LLM inference. Experiments were conducted using A100 GPUs equipped with 64 GB memory, allowing multiple parallel instances of the LLM per GPU. Specifically, we utilize the Qwen2.5-7B model and its quantized variant (Qwen2.5-awq), with the possibility to duplicate models within the same GPU to distribute the computational load.

As shown in Table 4, overhead naturally depends on backbone speed—the faster the backbone, the higher the relative overhead, as the LLM inference runs concurrently at a fixed speed. Nevertheless, by replicating the LLM across multiple GPUs, we reduce the overhead to virtually zero for all models. Specifically, using four LLM instances (two GPUs without quantization) eliminates overhead even for the fastest model. Moreover, quantization further reduces the computational requirement, allowing overhead elimination with just a single GPU hosting four quantized LLM replicas.

6 Pipeline overhead

Although SceneForge excludes the composed samples from the image–3D loss computation (applying it only on pre-rendered single-object views), image–3D retrieval performance remains largely unaffected.

Table 5 compares the performance of standard methods (OpenShape, ViT-Lens, and Uni3D) to their SceneForge-augmented counterparts (SF-variants) on Pix3D and Objaverse LVIS benchmarks.

These results highlight that SceneForge’s strategy of masking composed samples from the image–3D alignment does not degrade performance substantially. Specifically, while lower, accuracy variations remain within narrow margins across all three backbones, underscoring that SceneForge effectively preserves baseline performance while focusing primarily on enhancing text–3D alignment.

Model	Task	Top-k	1	2	3	4	5	6	7	8	9	10
OmniBind-F	Text→3D	Top-1	72.41	44.15	29.33	18.90	16.42	11.81	8.84	7.67	5.23	4.76
		Top-5	88.03	66.42	50.76	42.58	37.27	29.99	24.02	20.45	17.14	15.62
	3D→Text	Top-1	81.47	52.17	43.67	23.32	28.18	23.59	15.72	12.15	10.05	9.54
		Top-5	94.21	74.84	65.52	49.82	44.27	39.63	32.04	28.33	21.88	17.02
Uni3D	Text→3D	Top-1	71.64	42.59	28.54	19.88	15.10	12.12	8.70	6.76	5.74	4.63
		Top-5	89.16	68.72	54.66	43.63	36.46	31.22	25.24	20.91	18.42	15.73
	3D→Text	Top-1	79.90	48.54	34.27	24.94	18.92	14.87	10.52	7.55	6.01	4.76
		Top-5	92.71	74.60	61.93	51.58	43.37	37.23	29.45	23.30	19.85	16.59
SF-Uni3D (N=2)	Text→3D	Top-1	72.70	77.37	76.78	74.91	71.40	63.41	48.54	35.72	23.56	17.34
		Top-5	88.24	91.46	90.12	88.01	85.13	78.02	64.28	52.53	39.74	31.41
	3D→Text	Top-1	81.50	82.87	80.02	78.09	69.40	61.99	54.26	43.68	23.04	16.86
		Top-5	93.25	94.14	92.48	90.63	85.22	77.23	70.51	60.92	38.61	31.09
SF-Uni3D (N=3)	Text→3D	Top-1	73.61	74.37	78.80	77.77	73.97	66.94	53.06	39.99	29.01	20.67
		Top-5	86.24	90.41	92.20	91.65	89.60	85.07	75.67	63.67	51.84	40.78
	3D→Text	Top-1	84.40	80.74	84.62	83.02	81.98	78.06	64.70	48.85	35.43	24.26
		Top-5	91.01	93.00	93.94	93.99	93.21	91.74	85.24	74.89	63.10	50.30
SF-Uni3D (N=4)	Text→3D	Top-1	71.45	73.84	77.20	78.14	75.73	68.66	56.91	48.47	33.52	24.80
		Top-5	88.64	90.16	91.82	91.03	89.02	82.53	72.81	64.22	49.86	39.12
	3D→Text	Top-1	78.75	79.76	79.60	86.02	82.87	80.96	68.09	61.13	45.56	34.80
		Top-5	90.58	92.06	91.67	96.25	93.24	91.14	81.96	76.24	63.02	51.46
SF-Uni3D (N=5)	Text→3D	Top-1	71.02	72.73	77.54	77.76	78.16	69.95	58.13	50.71	37.46	28.43
		Top-5	87.09	89.18	92.21	92.45	92.86	83.54	73.62	64.75	52.63	40.92
	3D→Text	Top-1	78.58	79.07	81.26	82.64	85.24	86.45	71.47	68.41	59.34	36.57
		Top-5	90.29	91.52	93.65	95.23	96.82	97.18	85.53	82.06	74.44	50.75
OpenShape (N=1)	Text→3D	Top-1	63.04	37.48	25.11	17.49	13.28	10.67	7.65	5.95	5.05	4.07
		Top-5	82.92	63.91	50.83	40.58	33.90	29.03	23.48	19.44	17.13	14.63
	3D→Text	Top-1	70.31	42.71	30.16	21.94	16.65	13.09	9.26	6.64	5.29	4.19
		Top-5	86.22	69.38	57.60	47.97	40.33	34.62	27.39	21.67	18.46	15.43
SF-OpenShape (N=2)	Text→3D	Top-1	63.98	68.09	67.57	65.92	62.83	55.80	42.72	31.43	20.73	15.26
		Top-5	82.06	85.06	83.81	81.85	79.17	72.56	59.78	48.85	36.96	29.21
	3D→Text	Top-1	71.72	72.93	70.42	68.72	61.07	54.55	47.75	38.44	20.28	14.84
		Top-5	86.72	87.55	86.01	84.29	79.25	71.82	65.57	56.66	35.91	28.91
SF-OpenShape (N=3)	Text→3D	Top-1	64.78	65.45	69.34	68.44	65.09	58.91	46.69	35.19	25.53	18.19
		Top-5	80.20	84.08	85.75	85.23	83.33	79.12	70.37	59.21	48.21	37.93
	3D→Text	Top-1	74.27	71.05	74.07	72.26	72.14	68.69	56.94	43.00	31.18	21.35
		Top-5	84.09	86.49	87.36	87.41	86.68	85.32	79.27	69.65	58.68	46.78
SF-OpenShape (N=4)	Text→3D	Top-1	62.88	64.98	68.94	70.77	68.25	61.24	50.08	42.06	29.49	22.54
		Top-5	82.43	83.85	85.39	84.66	83.79	77.75	68.71	60.72	47.37	37.38
	3D→Text	Top-1	69.30	70.19	70.05	75.70	72.93	71.25	59.92	53.80	40.09	30.62
		Top-5	84.24	85.62	85.25	89.51	86.71	84.76	76.22	70.90	58.61	47.86
SF-OpenShape (N=5)	Text→3D	Top-1	62.50	64.00	68.24	68.43	68.78	61.56	51.15	44.62	32.96	24.02
		Top-5	80.99	82.94	85.76	85.98	86.36	77.69	68.47	60.22	48.95	38.06
	3D→Text	Top-1	69.15	69.56	71.51	72.72	74.21	76.08	62.89	60.20	52.22	32.18
		Top-5	83.46	84.61	87.09	88.56	90.04	90.38	77.54	74.32	67.23	45.20
ViT-Lens (N=1)	Text→3D	Top-1	67.34	40.03	26.82	18.70	14.19	11.39	8.17	6.34	5.38	4.31
		Top-5	85.60	66.97	53.47	42.89	35.00	29.97	24.23	20.06	17.74	15.09
	3D→Text	Top-1	75.11	45.61	32.81	23.84	18.13	14.00	9.89	7.09	5.63	4.48
		Top-5	88.00	71.62	59.45	49.52	41.63	35.74	28.27	22.37	19.10	16.09
SF-ViT-Lens (N=2)	Text→3D	Top-1	68.34	72.73	71.17	69.42	66.18	58.80	46.43	34.99	23.55	17.35
		Top-5	84.71	87.80	86.51	84.49	81.73	75.50	62.92	52.43	39.55	31.16
	3D→Text	Top-1	76.61	77.90	75.22	73.41	65.34	58.29	51.01	41.45	21.66	15.85
		Top-5	89.52	90.37	88.78	87.00	81.81	74.14	67.69	58.49	37.14	31.85
SF-ViT-Lens (N=3)	Text→3D	Top-1	69.19	69.91	74.07	73.10	69.53	62.92	49.88	37.59	27.27	19.43
		Top-5	82.79	86.79	88.51	87.98	86.02	81.67	72.65	61.12	49.77	39.15
	3D→Text	Top-1	79.34	75.89	79.54	78.04	77.06	73.38	60.02	45.92	33.30	22.81
		Top-5	87.37	89.28	90.18	90.23	89.49	88.06	81.84	71.90	60.58	48.29
SF-ViT-Lens (N=4)	Text→3D	Top-1	67.15	68.41	72.22	73.85	71.47	64.47	53.29	45.07	31.28	24.03
		Top-5	85.09	86.55	88.15	87.39	85.46	79.23	70.69	62.65	48.86	38.55
	3D→Text	Top-1	73.45	74.38	74.03	80.46	77.51	75.69	64.32	57.44	42.42	32.63
		Top-5	86.94	88.38	88.00	92.40	89.51	87.49	78.68	73.19	60.50	49.40
SF-ViT-Lens (N=5)	Text→3D	Top-1	66.76	68.36	72.89	72.99	73.87	65.75	54.64	47.66	35.21	26.72
		Top-5	83.64	85.61	88.53	88.75	89.15	80.20	70.67	62.16	50.02	38.48
	3D→Text	Top-1	73.87	74.33	76.38	77.68	80.13	81.26	67.18	64.31	55.22	34.37
		Top-5	86.09	87.13	89.91	91.42	92.95	93.29	82.05	78.49	71.45	48.72

Table 6: Extended Top-1 and Top-5 cross-modal retrieval accuracies (%) on the **N-LVIS** benchmark for $N = 1 \dots 10$. SF indicates SceneForge-trained variants.

7 Extended Results for N-Object Cross-Modal Retrieval

Table 6 reports full Top-1 and Top-5 retrieval accuracies on **N-LVIS** for $N=1 \dots 10$.¹ we consider the SF variants built on OpenShape, ViT-Lens and Uni3D back-bones, trained on different number of maximum composed objects $N \in \{1, \dots, 5\}$. This extended table allows three key observations:

1. **Compositional training generalizes across back-bones.** For *all* three backbones (Uni3D, ViT-Lens, OpenShape) the SF models that see multi-object scenes during training maintain high accuracy far beyond their training set complexity (e.g. SF-ViT-LENS $N=3$ keeps $\sim 63\%$ Top-1 at $N=6$, versus $< 30\%$ for the single-object ViT-Lens baseline).
2. **Each variant peaks near its training N .** The best performance for a given model is reached at or just beyond the composition size used in training, corroborating the trend discussed in the main paper. This ‘sweet-spot’ behaviour is consistent across back-bones, indicating that the phenomenon is architectural-agnostic.
3. **Single-object fidelity vs. multi-object robustness.** Models trained with $N>1$ incur only a modest drop at $N=1$ (e.g. SF-OPENSHAPE $N=5$ is 0.5 pp below its $N=1$ counterpart on text \rightarrow 3D Top-1), yet outperform their baselines by large margins at higher N . This emphasises the importance of compositional training when downstream applications require understanding crowded scenes.

Overall, the new results strengthen our main conclusion: scene-level composition is a backbone-agnostic strategy that substantially improves retrieval in complex multi-object settings while preserving competitive performance on standard single-object tasks.

8 Object Repositioning Training Details

We optimize a learnable spatial offset $\Delta = (\delta_x, \delta_y, \delta_z) \in \mathbb{R}^3$ to reposition the second object in a combined point cloud, aligning it with the updated caption while keeping the 3D-text encoder frozen. The objective is to maximize the cosine similarity between the modified point cloud embedding and the text embedding while regularizing the translation magnitude:

$$\mathcal{L} = -\cos(\phi(x^{3D} + \Delta), \psi(x^{txt})) + \lambda \|\Delta\|^2, \quad (4)$$

where $\phi(\cdot)$ and $\psi(\cdot)$ are the frozen embedding functions for 3D and text, respectively. The regularization term $\lambda \|\Delta\|^2$ prevents excessive displacement.

We optimize Δ using the Adam optimizer with a learning rate of 5×10^{-2} and weight decay of 10^{-4} . The regularization weight is set to $\lambda = 0.1$. Training runs for 200 iterations with a batch size of 1, as each optimization step refines a single point cloud-text pair. The offset Δ is initialized from a Gaussian distribution with mean 0 and standard deviation 0.01, ensuring small but nonzero initial displacements.

9 Analysis of Point Cloud Sampling Budget

In the main paper, we identified that zero-shot performance tends to degrade when composing a high number of objects ($N > 3$) under a fixed 10k point budget. To investigate the relationship between point resolution, scene complexity, and performance, we conducted a preliminary smaller-scale analysis which we detail here.

This experiment was run on a 50k random subset of the Objaverse dataset (without LVIS labels). We varied the point cloud resolution ($P \in \{5k, 10k, 15k, 20k\}$) and the maximum number of composed objects ($N \in \{1, 2, 3, 4, 5\}$), evaluating zero-shot top-1 accuracy on our standard suite of benchmarks. All other hyperparameters were kept fixed.

The consolidated results are presented in Table 7.

¹All models are trained on the full Objaverse-LVIS training split *excluding* the evaluation shapes, following the protocol we reported in the N-LVIS main work section.

Table 7: Zero-shot top-1 accuracy (%) on all benchmarks with varying point budgets and object counts. The best result for each point budget is highlighted.

Dataset	Points	N=1	N=2	N=3	N=4	N=5
LVIS	5k	27.24	27.59	27.26	27.02	26.58
	10k	27.41	28.06	28.15	27.64	27.33
	15k	27.38	28.03	28.18	28.09	27.79
	20k	27.47	28.10	28.13	28.13	28.10
ModelNet	5k	65.04	65.51	65.11	64.27	63.89
	10k	65.12	65.83	66.28	65.51	65.04
	15k	65.17	65.85	66.26	66.21	65.58
	20k	65.15	66.04	66.25	66.25	66.28
ScanObjNN	5k	48.98	49.53	48.80	48.52	47.96
	10k	49.03	49.71	50.01	49.45	49.08
	15k	49.10	49.80	50.04	50.11	49.68
	20k	49.08	49.92	50.09	50.05	50.09
ScanNet	5k	25.98	26.47	26.25	25.04	24.79
	10k	26.17	27.23	27.45	26.81	26.10
	15k	26.15	27.22	27.48	27.44	27.09
	20k	26.21	27.28	27.44	27.47	27.42

Two main conclusions can be drawn from these results. First, the benefit of compositions is evident at $N = 2$ and generally peaks or plateaus around $N = 3$. This aligns with the main paper’s finding that $N = 3$ is an optimal configuration for the 10k point budget. Second, a higher point budget primarily serves to alleviate performance degradation at high scene complexity ($N = 4, N = 5$) rather than dramatically increasing peak accuracy. For instance, at 20k points, the performance drop is nearly eliminated. This confirms that the degradation is not a flaw in the compositional method itself but a direct result of feature fragmentation under a constrained point budget.

References

- [1] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [2] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023.
- [3] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.
- [4] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL <https://huggingface.co/Qwen/Qwen2.5-VL-32B-Instruct-AWQ>. Qwen2.5 is licensed under the Apache 2.0 license.
- [5] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. In *International Conference on Learning Representations (ICLR)*, 2024. URL <https://github.com/baaivision/Uni3D>. This work is licensed under the MIT License.