ADA3D: ADAPTIVE CROSS-DIMENSIONAL ALIGNMENT FOR UNIVERSAL 3D REPRESENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Motivated by the significant performance of contrastive learning models in vision-language tasks, recent efforts have extended such architecture into 3D scenes for richer representation. However, existing approaches typically utilize the pretrained 2D ViT models to encode original 3D group features, which overlooks the inherent domain gap between these two modalities and leads to suboptimal performance. Observing that 3D point clouds and 2D images share geometric correspondences that can be explicitly correlated through cross-dimension projection, we propose Ada3D, a framework that focuses on the adaptive transfer from the 3D to 2D feature spaces to enable universal 3D representation. We achieve this transfer capability via a dual mechanism comprising a pre-adapter and a post-adapter. The pre-adapter bridges unordered point cloud groups and regular image patch sequences through the point-level and patch-level fusion paradigm. Subsequently, the post-adapter utilizes a mix-attention mechanism to capture the latent features in the refined 3D groups and transforms them into the ViT input space. Furthermore, we introduce a specialized Gaussian masked autoencoder and a shallow Transformer layer finetuning strategy to preserve the generalization and reduce computation. Extensive experiments demonstrate that our approach achieves remarkable results on diverse 3D representation tasks, including zero-shot open-vocabulary classification, segmentation, and visual question answering.

1 INTRODUCTION

General 3D representation has become a fundamental demand for various real-world applications, such as autonomous driving Huang et al. (2022); Ning et al. (2025; 2026), embodied intelligence Gupta et al. (2021), and augmented reality Azuma (1997). Meanwhile, many works with contrastive learning frameworks Radford et al. (2021); Jia et al. (2021); Sun et al. (2023); Li et al. (2023a) have demonstrated remarkable success in universal image-text representations. Benefiting from concise self-supervised pre-training on abundant web-scraped datasets, these works achieve promising performance across a range of upstream and downstream tasks Wang et al. (2025); Zhai et al. (2023); Yu et al. (2023). Therefore, it is both intuitive and practical to extend the contrastive learning paradigm to accommodate 3D sensory inputs (e.g., LiDAR point clouds). Previous works Liu et al. (2023b); Xue et al. (2023); Qi et al. (2023) typically employ conventional 3D encoders (e.g., PointNet Qi et al. (2017a) or SparseConv Choy et al. (2019)) to extract global features and directly align them with the `[CLS]` token of the image or the text generated by pretrained models, denoted as the "Pipeline1" in Figure 1. Though the results in zero-shot retrieval or classification tasks demonstrate the effectiveness of the above approaches, their universality remains constrained due to the structural rigidity of the vanilla point encoder.

Compared with group-based Qi et al. (2017a;b) or CNN-based Choy et al. (2019) encoders, the Transformer Vaswani et al. (2017) architecture exhibits stronger representational capacity, and the generalization capability of pretrained 2D ViTs in CLIP makes their integration into 3D encoders a natural choice. Therefore, recent works Zhu et al. (2023b); Zhang et al. (2022); Zhou et al. (2023a) attempt to directly inject 3D encoder outputs into a pretrained 2D ViT, illustrated as "Pipeline2" in Figure 1. However, these methods neglect the inherent discrepancy between the two feature spaces. Concretely, the 3D encoder employs farthest point sampling (FPS) to produce an irregular set of point groups, whereas the 2D ViT expects a regular $N \times N$ grid of image patches. Meanwhile, the geometric and structural information in 3D features is different from the patch representation in 2D

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

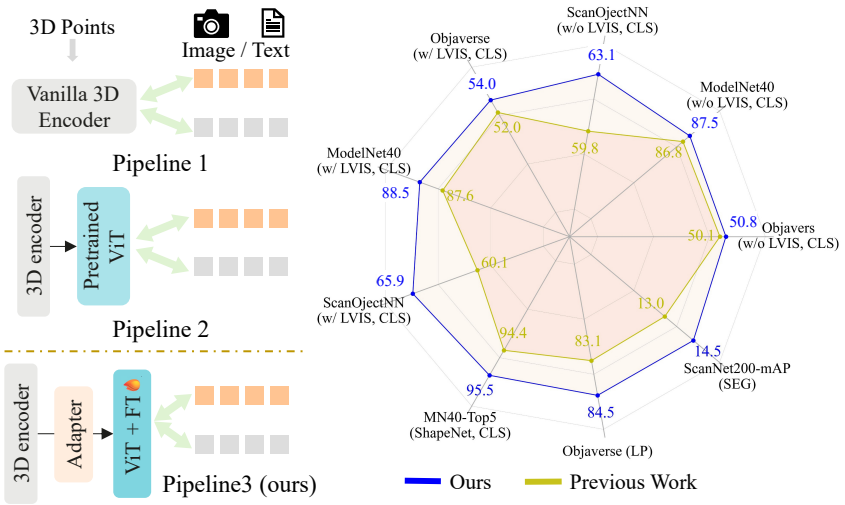


Figure 1: The comparison of different pipelines utilized for universal 3D understanding. Our method could achieve superior performance across diverse tasks with a specific adapter designed.

ViTs. Moreover, naively bridging this gap with a simple MLP or cross-attention mechanism yields only limited improvement compared with a direct connection, as evidenced by the 3D zero-shot classification performance in Table 1. Therefore, designing an effective adapter that reconciles both the structural and content discrepancies between these two spaces is crucial.

Table 1: Zero-shot Top-1 accuracy (%) with different adapters. We employ the PointNet Qi et al. (2017a) as 3D encoder and CLIP-bigG/14 Radford et al. (2021) as 2D ViT.

Adapter	OBJ-LVIS	MN40	ScanOBJNN
w/o	42.1	71.2	54.7
MLP	41.6	73.4	55.4
CrossAttn	45.1	81.1	60.9
Ours	**54.0**	**88.5**	**65.9**

In this paper, we present Ada3D, a novel framework introduced to bridge the feature spaces between the traditional 3D encoder and 2D ViT. Since point clouds and images can be inherently connected through cross-dimensional projection, we exploit this relationship to align their structural information effectively. First, we propose the pre-adapter to leverage the 2D geometric priors existing in the positional embeddings of a pretrained ViT, which consists of a multi-view aggregation (MVA) module and a patch content interaction (PCI) module. The MVA module projects the group points into multiple 2D views and aggregates the positional embeddings at the point level, while the PCI module facilitates patch-level interaction between the group features through a self-attention mechanism. Subsequently, we introduce the post-adapter, which initializes a series of queries and conducts a mix-attention mechanism to further capture the latent information and align with the embedding space. Moreover, we apply the Gaussian masked autoencoder (MAE) and shallow Transformer layer finetuning strategy to ensure the efficacy of the model under an affordable training burden. As shown on the right of Figure 1, our method exhibits superior performance over the existing methods on diverse 3D open-vocabulary tasks. This suggests that our approach is not only applicable to upstream scenarios such as zero-shot and few-shot classification but can also be adopted to the downstream segmentation task. Moreover, by integrating with visual language models (VLMs), our Ada3D can be extended to the visual question answering (VQA) field. Overall, the contribution of our paper can be summarized as:

- We propose the Ada3D, a novel framework focuses on the discrepancy between the point features and 2D ViT for universal 3D representation.
- The introduced pre-adapter fuses point-level and patch-level geometric priors with 3D group features, while the post-adapter further acquires the latent information and aligns with the ViT embedding space. Integrating with the Gaussian MAE and shallow layers finetuning strategy, our model effectively improves the representation capability.
- The superior performance in various upstream and downstream open-vocabulary 3D representation tasks demonstrates the fantastic application of our model.

2 Related Work

2.1 Vision-Language Pretraining Model

The advanced vision-language pretraining models, such as CLIP Radford et al. (2021), Flamingo Alayrac et al. (2022), BLIP Li et al. (2023a), and LiT Zhai et al. (2022), demonstrate that leveraging the self-supervised contrastive learning framework on the numerous internet-source datasets enables significant performance in the open-vocabulary classification and retrieval tasks Sun et al. (2024; 2023); Shi et al. (2025). The structurally similar encoding branches are applied to extract respective information in the image and text modality, followed by an alignment paradigm to convert them into a joint embedding space. Furthermore, EVA-CLIP Sun et al. (2023) conducts novel techniques in the representation and optimization for a stable and efficient training process. MaskCLIP Dong et al. (2023) and FLIP Li et al. (2023b) effectively reduce computational costs and enhance the encoding capability by masking a proportion of image patches. Benefiting from the generalization ability of the pretrained models, several works apply them to the downstream vision tasks, including segmentation Zhang et al. (2024a); Dong et al. (2023); Zhou et al. (2023b); Li et al. (2022a) and detection Siméoni et al. (2021); Gu et al. (2021); Li et al. (2022b). Our Ada3D aims to extend the powerful capability of the pretrained model to the 3D LiDAR representation, and align it with the shared image-text embedding space to enable various tasks in the real world.

2.2 Universal Representation for 3D Point Clouds

As a crucial component of scene understanding and spatial intelligence Yang et al. (2025); Chabris et al. (2006), 3D representation learning Wang et al. (2019); Yu et al. (2021); Qi et al. (2017a); Choy et al. (2019) has witnessed substantial progress in recent years. Early approaches Pang et al. (2023); Wang et al. (2021); Yu et al. (2022) rely only on the 3D data and utilize a reconstruction task for scene encoding. The absence of training on the multi-modal dataset hinders their applicability in the open-vocabulary tasks. To solve this issue, PointCLIP Zhang et al. (2022) introduces vision-language models and aligns point cloud features with image and text embeddings, achieving notable improvements in open-world understanding. Subsequently, ULIP2 Xue et al. (2024) and OpenShape Liu et al. (2023b) further construct large-scale 3D–image–text pairs dataset to boost performance. Despite these efforts, such methods are still limited by the lightweight 3D encoders compared with the Transformer architectures in CLIP. Therefore, ViT-Lens Lei et al. (2024) and Uni3D Zhou et al. (2023a) further adopt pretrained ViTs as a partition of the 3D encoding branch to scale the ability of models. However, they rigidly align the outputs of 3D encoders with ViT inputs without effectively mitigating the domain gap between unordered 3D point sets and structured 2D image patches in the feature space. To address this issue, our work designs an effective strategy to bridge this modality gap, excavating the full potential of ViT for universal 3D representation learning.

2.3 Multimodal Foundation Model

Recently, large language models (LLMs) Touvron et al. (2023); Chiang et al. (2023); Floridi & Chiriatti (2020) have achieved remarkable progress across a broad spectrum of understanding and generating tasks. Inspired by their powerful decoding capabilities, substantial efforts Liu et al. (2024); Zhu et al. (2023a) integrate LLMs into the visual encoder and construct a series of multimodal foundation models (MFMs) such as BLIP2 Li et al. (2023a), LLaVA Liu et al. (2023a), and Flamingo Alayrac et al. (2022). These methods typically employ a lightweight projector to correlate the visual encoder with the LLM, which is trained via a two-stage finetuning strategy. By leveraging the ability to interpret visual tokens, such MFMs can be extended to visual question answering (VQA) Dai et al. (2023a) and text-to-image (T2I) generation Rombach et al. (2022) tasks. Notably, the ViT encoder remains frozen in such paradigms, which ensures the consistent visual representation across various models. In this work, we aim to construct unified multimodal representations such that the output from our 3D branch can be seamlessly incorporated into existing MFMs to facilitate 3D understanding and generation capabilities.

3 Method

As illustrated in Figure 2, our Ada3D consists of three branches: an image branch, a text branch, and a core 3D branch in our framework. Notably, the 3D branch contains a standard point-based encoder (termed as 'point tokenizer' hereafter), the proposed pre-adapter and post-adapter, followed by a

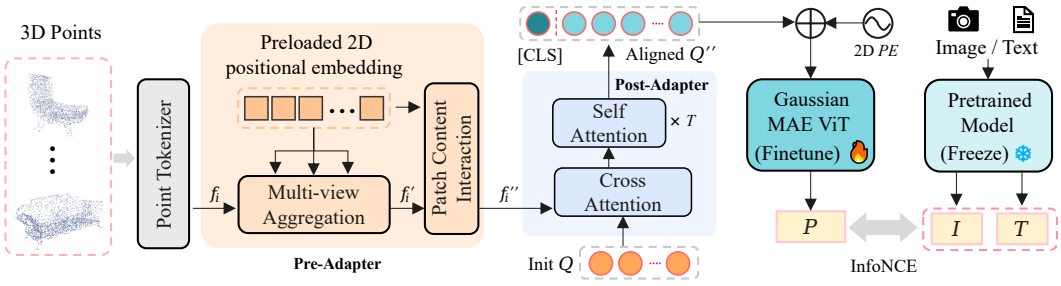


Figure 2: Overview of Ada3D model. Point features are first extracted via a normal point tokenizer. A pre-adapter performs cross-dimension feature fusion, while a post-adapter aligns latent features with the ViT input space. Gaussian MAE and shallow-layer finetuning are applied to encode the feature.

pretrained 2D ViT model. Furthermore, the overall architecture is trained via a contrastive learning paradigm within a Gaussian masked autoencoder (MAE) and shallow layer finetuning strategy.

3.1 Pre-Adapter for Point- and Patch-Level Fusion

Multi-View Aggregation (MVA) Module. Given the irregular 3D input points, we first feed them into the point tokenizer, which consists of a fastest point sampling (FPS) and a K-nearest neighbors (KNN), followed by a PointBERT Yu et al. (2022). These raw 3D points are tokenized as the group points set $\{(f_i, c_i), i \in [1, M]\}$, where M defines the number of groups, $f_i \in \mathbb{R}^D$ denotes the group feature, and $c_i \in \mathbb{R}^3$ represents the corresponding 3D center coordinate of each group. To inject 2D geometric priors into these 3D groups, we leverage the positional embedding (termed as PE) from the pretrained ViT and incorporate it into the group features, as illustrated in Figure 3 (a). Specifically, we project 3D center coordinate c_i into V orthorhombic perspectives to obtain its corresponding 2D projections $\{c_i^j | j \in [1, V]\}$, where $c_i^j \in \mathbb{R}^2$ denotes the projected coordinates under the j-th view. Then, the 2D projection map is divided into $K \times K$ patches, whose resolution matches that of the positional embeddings in the pretrained model. In each point-level positional embedding (PPE) fusion block, we first perform bilinear interpolation on the positional embedding map to sample the embedding value for each projected point, and then apply an MLP layer to align the channel dimension with the original features. The process is defined as follows:

$$f_i^j = \text{MLP}(\text{BilInter}(PE, [\frac{c_i^j[0]}{x_j}, \frac{c_i^j[1]}{y_j}])), \tag{1}$$

where x_j and y_j denote each patch size in the x- and y-axes of the j-th view, respectively. After obtaining the interpolated embedding value f_i^j from all views, we conduct an averaging operation for them and fuse the result with the original group feature via an addition operation to yield f_i'.

Patch Content Interaction (PCI) Module. After the MVA module, the enhanced group feature f_i' acquires explicit geometric awareness but remains inconsistent with the patch-based structure. To solve the problem, we introduce the PCI module. As illustrated in Figure 3 (b), given a patch $p_n (n \in [1, K^2])$ on the 2D map, we aggregate all group features that are projected within the current patch. These features are first passed through a max-pooling operation, followed by a batch normalization and a ReLU activation, to form each patch representation f_{p_n}. Subsequently, we add the PE with the patch feature sequence $F_p = [f_{p_1}, \cdots, f_{p_{K^2}}]$ and feed them into the Transformer block, which can be formulated as:

$$\begin{aligned} [Q_p, K_p, V_p] &= (F_p + PE) \cdot [W_Q^T, W_K^T, W_V^T], \\ F_p' &= \text{Softmax}(\frac{Q_p K_p^T}{\sqrt{D}} \cdot \mathbf{M}_p) V_p; \mathbf{M}_p = I_p I_p^T, \end{aligned} \tag{2}$$

where $\mathbf{M}_p$ denotes the attention mask and I_p is an indicator function specifying whether the current patch contains projected points. The mask is designed to ensure only non-empty patches are allowed to interact with each other, thereby reducing computational redundancy and mitigating the interference from invalid patches. The Transformer block is replicated three times to enable effective patch-level

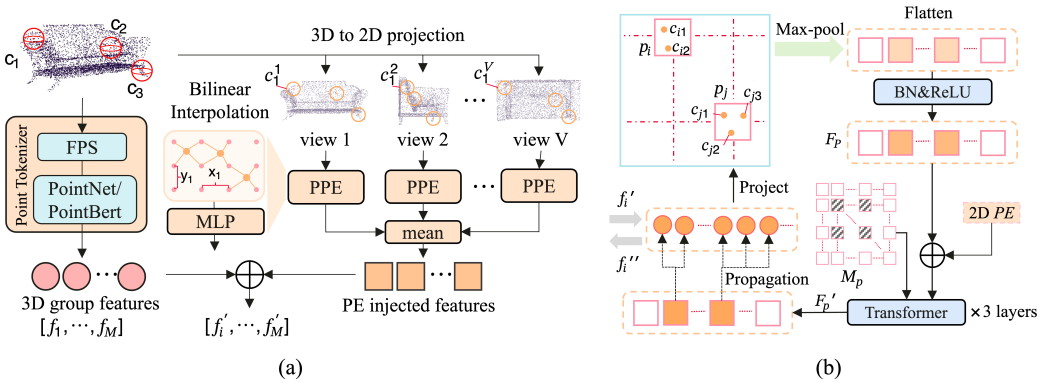


Figure 3: Illustration of the MVA and PCI modules. (a) MVA projects each group's 3D center onto orthorhombic 2D views, samples ViT positional embeddings via bilinear interpolation, and fuses them into group features. (b) PCI aggregates group features into patch-level representations, refines them with masked Transformer blocks, and propagates the updates back.

interactions. Finally, the updated patch features are propagated back to their corresponding group features based on the projection coordinates, yielding the refined representation f_i''. Overall, the PCI module injects patch-level structural awareness into the 3D group points and achieves a refined alignment between the 3D tokenizers and the ViT input sequence.

3.2 Mix-Attention Mechanism in Post-Adapter

After obtaining the point-level and patch-level enhanced group features f_i'', we further employ a post-adapter to convert the group features space into the 2D tokens space before feeding them into the pretrained ViT. Specifically, we initialize a set of learnable queries $Q \in \mathbb{R}^{M\times D}$ and apply a cross-attention layer to excavate the latent information in the 3D representation, yielding latent queries Q'. Subsequently, we conduct a self-attention tower to adapt the Q' into the ViT input. The overall process in the mix-attention mechanism can be defined as:

$$\begin{aligned} Q' &= \text{CrossAttn}(Q, [f_1'', \cdots, f_M'']), \\ Q'' &= [\text{SelfAttn}(Q')]_{\times T}, \end{aligned} \tag{3}$$

where Q'' is the output of our post-adapter, which will be combined with an initialized `[CLS]` token and fed into the following ViT model for the 3D representation generation.

3.3 Finetuning Strategy and Training Objective

During training, the modules preceding the ViT backbone (i.e., the 3D tokenizer, pre-adapter, and post-adapter) are trained from scratch, while the ViT is trained with a Gaussian MAE and shallow layer finetuning strategy. Specifically, instead of replacing masked tokens with zero-value expanding or dropping them entirely like previous works He et al. (2022); Pang et al. (2023), we sample a ratio of tokens and multiply them with Gaussian modulation parameters to enhance the model's robustness and prevent feature collapse. The distribution of the modulation parameters follows $\mathcal{N}(\mu = 0, \delta = 1)$. Hard masking removes tokens and introduces structural discontinuities, which often exacerbates the already irregular distribution of 3D tokens. In contrast, our Gaussian modulation preserves the complete token sequence and acts as a regularizer against domain-specific noise, thereby facilitating robust 3D-to-2D adaptation.

Furthermore, motivated by the observation in prior studies Jaegle et al. (2021) that deeper Transformer layers in a ViT show limited feature diversity, whereas the shallow layers exhibit more significant differences, we unfreeze only the first two shallow Transformer layers during finetuning to improve alignment efficiency while preserving the generalization ability.

Our overall training objective is to enforce cross-modal consistency within a unified embedding space, which is achieved by minimizing the information noise-contrastive estimation (InfoNCE) loss as:

$$\mathcal{L} = -\frac{1}{4N}\sum_{i=1}^{N}(\log\frac{\exp(P_i \cdot I_i/\tau)}{\sum_j \exp(P_i \cdot I_j/\tau)} + \log\frac{\exp(I_i \cdot P_i/\tau)}{\sum_j \exp(I_i \cdot P_j/\tau)} + \log\frac{\exp(P_i \cdot T_i/\tau)}{\sum_j \exp(P_i \cdot T_j/\tau)} + \log\frac{\exp(T_i \cdot P_i/\tau)}{\sum_j \exp(T_i \cdot P_j/\tau)}), \tag{4}$$

where P_i, I_i, and T_i represent the normalized 3D points, 2D image, and text features. N is the total pairs in a batch and τ refers to the learnable temperature.

4 Experiments

4.1 Experimental Setup

Datasets for 3D Representation Learning. We conduct contrastive learning pretraining on object-level 3D–image–text pairs datasets. Following previous approaches Liu et al. (2023b); Zhou et al. (2023a); Lei et al. (2024), our ensembled training dataset comprises ShapeNet (52.5k samples) Chang et al. (2015), 3D-FUTURE (16.6k) Fu et al. (2021), ABO (8.0k) Collins et al. (2022), and Objaverse (798.8k)Deitke et al. (2023). Additionally, we incorporate rendered images and textual descriptions from prior studies Lei et al. (2024). For open-vocabulary classification evaluation, we assess our model on three representative benchmarks: ModelNet40 Wu et al. (2015), ScanObjectNN Uy et al. (2019), and Objaverse-LVISDeitke et al. (2023). ModelNet40 and ScanObjectNN are widely used 3D classification datasets, where the former comprises 40 categories with approximately 12K computation-assigned design (CAD) samples, and the latter includes 15 common categories with around 15K real-world object scans. Objaverse-LVIS serves as a curated and refined subset of Objaverse Deitke et al. (2023), containing 47K high-quality instances spanning 1,156 LVIS categories.

Details of Implementation. Our proposed Ada3D is trained on $16\times$ A800 (80G) GPUs. The batch size per GPU varies according to the scale of the pretrained ViT backbone, which is set to 32 for ViT-G/14 and 64 for both ViT-L/14 and ViT-B/16. We adopt the Adam optimizer with a learning rate of 2e-4, and train the model for 200 epochs in total. The image and text branches are loaded from the Open-CLIP Radford et al. (2021). To mitigate computational redundancy and improve training efficiency, the visual and textual embeddings are precomputed and cached before the 3D pretraining stage. In the 3D branch, we partition each 3D scene into 512 groups. The channel dimension of the resulting 3D embeddings is carefully aligned with that of the positional embeddings, enabling the seamless feature fusion within the pre-adapter. In the post-adapter, the number of initial queries is also set to 512, maintaining a correspondence with the 3D tokens. In evaluation, we uniformly sample 10,000 colored points for each instance on Objaverse-LVIS Deitke et al. (2023), 10,000 uncolored points on ModelNet40 Wu et al. (2015), and 8,192 uncolored points on ScanObjectNN Uy et al. (2019).

4.2 Zero-shot 3D Classification

Zero-shot classification is a fundamental task in universal 3D representation learning and serves as a key indicator of a model's global perception capability. We train Ada3D-G on both the "NO LVIS" subset (excluding all shapes belonging to Objaverse-LVIS) and the full ensembled dataset, following the same evaluation protocol as OpenShape Liu et al. (2023b). We evaluate Top-1/3/5 accuracy on Objaverse-LVIS Deitke et al. (2023), ModelNet40 Wu et al. (2015), and ScanObjectNN Uy et al. (2019). As shown in Table 2, our model comprehensively surpasses prior methods, including ViT-Lens Lei et al. (2024) and Uni3D Zhou et al. (2023a) on the "NO LVIS" subset. When trained on the full dataset, Ada3D further achieves Top-1 accuracy of 54.0%, 88.5%, and 65.9% on the three benchmarks, surpassing all previous state-of-the-art approaches.

4.3 Regional Component Retrieval

Beyond evaluating global representations, we further analyze the model's sensitivity to regional information across diverse object point clouds. Specifically, we extract all token features from the ViT architecture within the 3D branch and project them into the final embedding space via the visual

Table 2: Zero-shot 3D classification on Objaverse-LVIS Deitke et al. (2023), ModelNet40 Wu et al. (2015) and ScanObjectNN Uy et al. (2019) dataset. "Ensembled" Liu et al. (2023b): pretraining with four datasets, Objaverse Deitke et al. (2023), ShapeNet Chang et al. (2015), ABO Collins et al. (2022) and 3D-FUTURE Fu et al. (2021).

Method	Training Data	Objaverse-LVIS			ModelNet40			ScanObjectNN		
		Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
PointCLIP Zhang et al. (2022)	w/o 3D data	1.9	4.1	5.8	19.3	28.6	34.8	10.5	20.8	30.6
PointCLIPv2 Zhu et al. (2023b)		4.7	9.5	12.9	63.6	77.9	85.0	42.2	63.3	74.5
ULIP-2 Xue et al. (2024)	Ensembled (no LVIS)	21.4	38.1	46.0	71.4	84.4	89.2	46.0	66.1	76.4
OpenShape-SparseConv Liu et al. (2023b)		37.0	58.4	66.9	82.6	95.0	97.5	54.9	76.8	87.0
OpenShape-PointBERT Liu et al. (2023b)		39.1	60.8	68.9	85.3	96.2	97.4	47.2	72.4	84.7
ViT-Lens Lei et al. (2024)		50.1	71.3	78.1	86.8	96.8	97.8	59.8	79.3	87.7
Ada3D		**50.8**	**72.3**	**79.1**	**87.5**	**97.6**	**98.8**	**63.1**	**83.7**	**91.0**
ULIP-2 Xue et al. (2024)	Ensembled	26.8	44.8	52.6	75.1	88.1	93.2	51.6	72.5	82.3
OpenShape-SparseConv Liu et al. (2023b)		43.4	64.8	72.4	83.4	95.6	97.8	56.7	78.9	88.6
OpenShape-PointBERT Liu et al. (2023b)		46.8	69.1	77.0	84.4	96.5	98.0	52.2	79.7	88.7
TAMM Zhang et al. (2024b)		50.7	73.2	80.6	85.0	96.6	98.1	55.7	80.7	88.9
ViT-Lens Lei et al. (2024)		52.0	73.3	79.9	87.6	96.6	98.4	60.1	81.0	90.3
OpenDlign Mao et al. (2024)		–	–	–	86.2	96.6	97.5	60.5	82.5	90.4
RECON++ Qi et al. (2024)		53.7	75.8	82.0	87.3	95.4	96.1	65.4	84.1	89.7
Uni3D Zhou et al. (2023a)		53.5	75.5	82.0	87.3	**98.1**	**99.2**	63.9	84.9	**91.7**
Ada3D		**54.0**	**75.9**	**82.9**	**88.5**	97.1	**99.2**	**65.9**	**85.1**	**91.7**

projection layer. This enables the calculation of a similarity score between each 3D token and the text embeddings, thereby assessing the model's capability to associate textural descriptions with specific local regions (e.g., *"the backrest of a chair"* or *"the legs of something or somebody"*). As illustrated in Figure 4, we visualize the comparison between our method and OpenShape-PointBERT Liu et al. (2023b). The red-highlighted points indicate the center of the top-20 3D tokens that exhibit high similarity to the given text. The results prove that our model can accurately localize the components that are aligned with the input description, showing that our approach not only encodes global representation within the `[CLS]` token but also preserves regional information in other tokens.

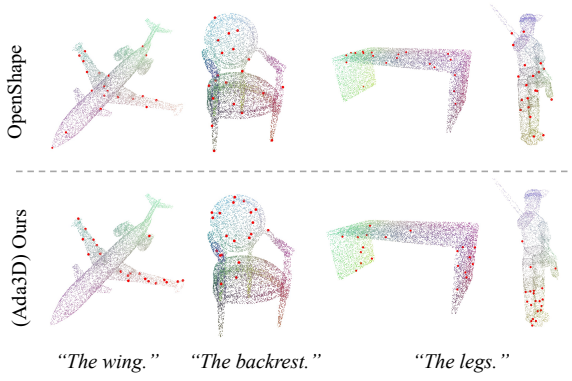


Figure 4: Comparison of the regional component recognized by Ada3D and previous work OpenShape Liu et al. (2023b). The marked red points indicate the top-20 group centers with the highest similarity scores to the input texts.

4.4 Open-Vocabulary Segmentation

We also evaluate whether Ada3D transfers to open-vocabulary 3D segmentation. Following prior works Yang et al. (2024); Lee et al. (2025), we load the pretrained point tokenizer weights and finetune it with a lightweight segmentation head on ScanNet200 Dai et al. (2017), preserving open-vocabulary generalization while adapting to dense prediction. As shown in Table 3, Ada3D achieves state-of-the-art performance on both semantic and instance segmentation. For semantic segmentation, we obtain 14.5% f-mIoU and 27.5% f-mAcc, surpassing Mosaic3D Lee et al. (2025) by 1.5% and 3.0%, respectively. On instance segmentation, our model also shows consistent advantages across all AP thresholds. These results confirm that Ada3D effectively captures fine-grained semantic information transferable to dense prediction tasks.

4.5 Multi-Modal Question Answering

Benefiting from the consistent embedding space shared across the 3D point, image, and text modalities, our learned 3D representations can be directly integrated into existing multi-modal foundations models (MFMs), thereby enabling a broader application of multi-modal understanding and generation tasks. Taking visual question answering (VQA) as an example, we follow the InstructBLIP Dai et al. (2023b) paradigm by freezing both the abstractor and the LLM, while feeding the output features from either

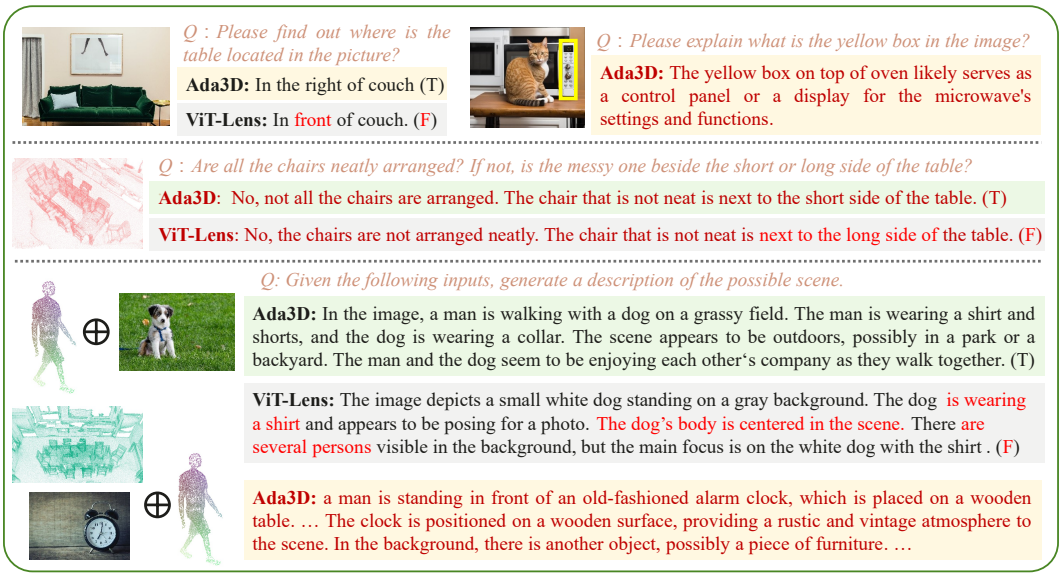


Figure 5: Examples of the visual question answering (VQA) employed on our Ada3D. The first and second rows represent the image-only and 3D-only inputs, respectively. The third row defines the multi-modal combined inputs. It can be observed that our model can recognize the accurate spatial relationship in the real-world scene.

Table 3: Comparison of the open-vocabulary segmentation task on ScanNet200 Dai et al. (2017). "Semantic" indicates the semantic segmentation, while the "Instance" denotes the instance segmentation.

Methods	Semantic		Instance		
	f-mIoU	f-mAcc	mAP	mAP/50	mAP/25
PLA Ding et al. (2023)	1.8	3.1	–	–	–
OpenScene-3D Peng et al. (2023)	6.4	12.2	4.8	6.2	7.2
RegionPLC Yang et al. (2024)	9.1	17.3	6.3	8.6	9.7
OpenIns3D Huang et al. (2024)	–	–	8.8	10.3	14.4
Mosaic3D Lee et al. (2025)	13.0	24.5	11.8	17.8	16.0
Ada3D	**14.5**	**27.5**	**12.2**	**18.2**	**16.2**

our 3D or 2D branch into the MFM. By incorporating the appropriate prompts, our model is capable of performing VQA in a training-free manner. As visualized in Figure 5, our framework flexibly supports multiple input configurations, including image-only, point-only, and hybrid image–point modalities. Notably, our model demonstrates a strong capability in spatial reasoning questions (e.g., "Where is the `[OBJ]` located?"), a competence that was largely absent in previous methods, such as ViT-Lens Lei et al. (2024). We attribute this improvement to the incorporation of geometric priors during pretraining, which enhances the model's spatial awareness and geometric consistency.

4.6 Ablation

Impact of Individual Modules. In this section, we conduct ablation studies to investigate the effectiveness of each component, using ViT-B/16 as the default backbone. As shown in Table 4, directly feeding the 3D tokenizer output into the pretrained ViT without any adapter yields only 34.09% Top-1 accuracy on Objaverse. Adding the MVA module to inject geometric priors brings a noticeable improvement, and further incorporating the PCI module for patch-level fusion raises Top-1 accuracy to 38.01%, a gain of approximately 4% over the vanilla framework. The post-adapter further boosts performance to 40% via mix-attention. Combined with MAE-based finetuning and shallow-layer unfreezing, Ada3D achieves 41.1%, 62.7%, and 70.6% in Top-1/3/5 accuracy, respectively. We additionally report trainable parameters and inference time, showing that our modules introduce only marginal computational overhead.

Table 4: Impact of different components proposed in our work. Additionally, the effectiveness of the Gaussian MAE and shallow layers finetuning strategy is also evaluated. We further report the trainable parameters and inference time for each configuration.

MVA	PCI	Post-Ada	+ MAE FT	Top1	Top3	Top5	Param. (M)	Time (ms)
			w/o	34.09	54.73	63.04	15.2	16.1
✓			w/o	36.90	56.91	65.08	16.4	16.5
✓	✓		w/o	38.01	58.46	66.43	19.7	17.9
✓		✓	w/o	39.79	59.01	67.98	20.2	18.5
✓	✓	✓	w/o	40.02	60.41	68.21	23.5	20.0
✓	✓	✓	w/	**41.07**	**62.68**	**70.59**	23.5	20.0

Table 5: Results of accuracy and parameter amounts under different finetuning strategies and the model scales.

	OBJ-LVIS		ScanOBJNN		Param.
	Top1	Top5	Top1	Top5	
Frozen	34.3	65.1	51.3	86.1	13M
LoRA FT (r = 8)	36.4	65.3	52.1	86.2	14.7M
FT top2 layers	**39.6**	**68.8**	**52.9**	**87.0**	21.3M
Ada3D-B/16	41.1	70.6	56.0	89.1	23.5M
Ada3D-L/14	44.9	73.9	53.6	87.9	92.1M
Ada3D-G/14	**50.8**	**79.1**	**63.1**	**91.0**	387M

Table 6: Comparison of different mask ratios and mask types in the MAE. Accuracy on the ModelNet40 Wu et al. (2015) is reported.

Ratio	Mask Type	Top1	Top3	Top5
0.0	/	68.4	86.1	95.1
0.15	Zero	69.2	87.5	95.3
0.25	Zero	70.4	88.7	95.3
0.25	Gaussian	**71.3**	**89.7**	**95.5**
0.45	Gaussian	68.3	86.3	95.3

Ablations on Finetuning Strategies and Model Scales. Table 5 presents the ablation on finetuning strategies. Freezing the entire ViT and training only the tokenizer and adapters (13M parameters) yields relatively low performance (34.31% and 51.31% Top-1 on Objaverse-LVIS and ScanObjectNN). We then explore LoRA-based finetuning Hu et al. (2022), showing that careful configuration design is crucial for 2D-to-3D transfer. Finetuning only the shallow Transformer layers further improves over LoRA, achieving 39.6% and 52.9% Top-1 accuracy with 21.3M trainable parameters.

We also investigate model scale using OpenCLIP-B/16, L/14, and bigG/14, all pretrained on LAION-2B Schuhmann et al. (2022). As shown in the lower part of Table 5, increasing capacity consistently improves accuracy: scaling from Ada3D-B to Ada3D-L yields a 3.8% Top-1 gain on Objaverse-LVIS, and further scaling to Ada3D-G adds another 5.9%, confirming that larger backbones provide richer 2D priors for cross-dimensional alignment. Our final Ada3D-G achieves 50.8% and 63.1% Top-1 on Objaverse-LVIS and ScanObjectNN, respectively.

Ablation on the Gaussian Masked Autoencoder In this section, we conduct an ablation study on the mask type and ratio within the pipeline. In detail, we explore two masking types: "Zero" indicates that the selected features are directly replaced with zeros, and "Gaussian" refers to the masked features being modulated via the Gaussian distribution parameter. As summarized in Table 6, the model attains a Top-1 accuracy of 68.% on the ModelNet40 dataset without applying any MAE. Employing a zero mask strategy with a ratio of 0.15 or 0.25 leads to a certain improvement over the baseline. Further substituting the mask type with the Gaussian modulation acquires an additional 0.9% gain. However, the accuracy drops markedly to 68.3% when the mask ratio increases to 0.45. We attribute this degradation to excessive masking, which leads to catastrophic information loss in point cloud representations. Consequently, we adopt the Gaussian mask strategy with a ratio of 0.25 as the default configuration in the final pretraining pipeline.

5 Conclusion

In this work, we introduced Ada3D, a novel framework bridging the intrinsic feature gap between 3D encoders and 2D ViTs. Using a pair of pre- and post-adapters, our model aligns unordered point groups with ordered image patches and refines their representations via mix-attention. In addition, the Gaussian MAE pretraining and shallow Transformer finetuning strategy enhance efficiency and stability. Our model yields state-of-the-art zero-shot 3D classification accuracy on Objaverse-LVIS, ModelNet40, and ScanObjectNN, and exhibits strong transferability to open-vocabulary segmentation and multi-modal question answering.

AI USE STATEMENT

In this work, we used generative AI tools to help polish the content of the paper and correct grammatical errors. All generated content was carefully checked by us to ensure that no hallucinations or errors were introduced.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Ronald T Azuma. A survey of augmented reality. *Presence: teleoperators & virtual environments*, 6 (4):355–385, 1997.

Christopher F Chabris, Thomas Jerde, Anita Williams Woolley, Margaret E Gerbasi, Jonathon P Schuldt, Jonathan Wai, Sean L Bennett, J Richard Hackman, and Stephen M Kosslyn. Spatial and object visualization cognitive styles: Validation studies in 3839 individuals. 2006.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3075–3084, 2019.

Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21126–21136, 2022.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023a.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023b.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13142–13153, 2023.

Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7010–7019, 2023.

Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10995–11005, 2023.

Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and machines*, 30(4):681–694, 2020.

Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129(12): 3313–3337, 2021.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei. Embodied intelligence via learning and evolution. *Nature communications*, 12(1):5721, 2021.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Keli Huang, Botian Shi, Xiang Li, Xin Li, Siyuan Huang, and Yikang Li. Multi-modal sensor fusion for auto driving perception: A survey. *arXiv preprint arXiv:2202.02703*, 2022.

Zhening Huang, Xiaoyang Wu, Xi Chen, Hengshuang Zhao, Lei Zhu, and Joan Lasenby. Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation. In *European Conference on Computer Vision*, pp. 169–185. Springer, 2024.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pp. 4651–4664. PMLR, 2021.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Junha Lee, Chunghyun Park, Jaesung Choe, Yu-Chiang Frank Wang, Jan Kautz, Minsu Cho, and Chris Choy. Mosaic3d: Foundation dataset and model for open-vocabulary 3d segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14089–14101, 2025.

Weixian Lei, Yixiao Ge, Kun Yi, Jianfeng Zhang, Difei Gao, Dylan Sun, Yuying Ge, Ying Shan, and Mike Zheng Shou. Vit-lens: Towards omni-modal representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26647–26657, 2024.

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022a.

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*, pp. 19730–19742. PMLR, 2023a.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975, 2022b.

Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23390–23400, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023a.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024.

Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. *Advances in neural information processing systems*, 36:44860–44879, 2023b.

Ye Mao, Junpeng Jing, and Krystian Mikolajczyk. Opendlign: Open-world point cloud understanding with depth-aligned images. *Advances in Neural Information Processing Systems*, 37:101144–101167, 2024.

Zhiwei Ning, Zhaojiang Liu, Xuanang Gao, Yifan Zuo, Jie Yang, Yuming Fang, and Wei Liu. Cmf-iou: Multi-stage cross-modal fusion 3d object detection with iou joint prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Zhiwei Ning, Xuanang Gao, Jiaxi Cao, Runze Yang, Huiying Xu, Xinzhong Zhu, Jie Yang, and Wei Liu. Fore-mamba3d: Mamba-based foreground-enhanced encoding for 3d object detection. *arXiv preprint arXiv:2602.19536*, 2026.

Yatian Pang, Eng Hock Francis Tay, Li Yuan, and Zhenghua Chen. Masked autoencoders for 3d point cloud self-supervised learning. *World Scientific Annual Review of Artificial Intelligence*, 1: 2440001, 2023.

Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 815–824, 2023.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017a.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017b.

Zekun Qi, Runpei Dong, Guofan Fan, Zheng Ge, Xiangyu Zhang, Kaisheng Ma, and Li Yi. Contrast with reconstruct: Contrastive 3d representation learning guided by generative pretraining. In *International Conference on Machine Learning*, pp. 28223–28243. PMLR, 2023.

Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. Shapellm: Universal 3d object understanding for embodied interaction. In *European Conference on Computer Vision*, pp. 214–238. Springer, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Bowen Shi, Peisen Zhao, Zichen Wang, Yuhang Zhang, Yaoming Wang, Jin Li, Wenrui Dai, Junni Zou, Hongkai Xiong, Qi Tian, et al. Umg-clip: A unified multi-granularity vision generalist for open-world understanding. In *European Conference on Computer Vision*, pp. 259–277. Springer, 2025.

Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. *arXiv preprint arXiv:2109.14279*, 2021.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13019–13029, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1588–1597, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Bingchao Wang, Zhiwei Ning, Jianyu Ding, Xuanang Gao, Yin Li, Dongsheng Jiang, Jie Yang, and Wei Liu. Fix-clip: Dual-branch hierarchical contrastive learning via synthetic captions for better understanding of long text. *arXiv preprint arXiv:2507.10095*, 2025.

Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matt J Kusner. Unsupervised point cloud pre-training via occlusion completion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9782–9792, 2021.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5): 1–12, 2019.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1912–1920, 2015.

Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1179–1189, 2023.

Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, et al. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27091–27101, 2024.

Jihan Yang, Runyu Ding, Weipeng Deng, Zhe Wang, and Xiaojuan Qi. Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19823–19832, 2024.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025.

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36:32215–32234, 2023.

Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, and Jie Zhou. Pointr: Diverse point cloud completion with geometry-aware transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12498–12507, 2021.

Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19313–19322, 2022.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18123–18133, 2022.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8552–8562, 2022.

Yi Zhang, Meng-Hao Guo, Miao Wang, and Shi-Min Hu. Exploring regional clues in clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3270–3280, 2024a.

Zhihao Zhang, Shengcao Cao, and Yu-Xiong Wang. Tamm: Triadapter multi-modal learning for 3d shape understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21413–21423, 2024b.

Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. *arXiv preprint arXiv:2310.06773*, 2023a.

Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11175–11185, 2023b.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023a.

Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2639–2650, 2023b.

A PARAMETER DETAILS IN DIFFERENT SCALE MODELS.

In this section, we present a detailed comparison of the hyperparameters and training configurations used across different model scales, including Ada3D-Base, Ada3D-Large, and Ada3D-Giant. As summarized in Table 7, these variants mainly differ in backbone capacity, tokenization dimensions, the depth of the 3D ViT, and the total number of learnable parameters. All models share consistent optimizer settings and data augmentation strategies, while the larger variants employ higher-dimensional token embeddings and deeper architectures to enhance representational power. This systematic scaling ensures a fair comparison and further demonstrates the effectiveness and generality of our design under varying computational budgets.

Table 7: Comparison of the hyperparameters and experimental configuration in our different models.

	Ada3D-Base	Ada3D-Large	Ada3D-Gaint
Pretrained Model	ViT-B-16	ViT-Large-14	ViT-bigG-14
Group Number		512	
3D Tokenize Channel	768	1024	1664
SA Layers in Post-Adapter	2	2	4
Layers in 3D ViT	12	24	24
Finetuned Shallow Layers	2	2	4
Output Channel	768	1024	1664
Optimizer		AdamW	
Weight Decay		0.2	
Peak LR	2e-4	2e-4	1e-4
Batch Size / GPU	64	64	16
Epochs	200	200	100
3D Points Augmentation		RandScale / RandShift / RandRotate	
Learnable Parameters	23.5M	92.1M	347M

B TRAINING ANALYSIS.

Based on the proposed pre-adapter and post-adapter designs, our method not only achieves superior performance but also exhibits significantly faster convergence during training. Taking ViT-B-16 as an example, we compare our approach with the ViT-Lens baseline in terms of Top-1 accuracy on Objaverse and ModelNet40 across different training epochs, as depicted in Figure 6. Under identical configurations of batch size and learning rate, our model converges within approximately 50 epochs and attains comparable performance to ViT-Lens, which needs about 100 epochs to achieve the same accuracy, indicating a significantly faster convergence behavior. This convergence discrepancy becomes even more pronounced as the model scales increase. We attribute this advantage to the cross-dimension features fusion introduced by our pre-adapter, which effectively alleviates the transfer burden of the post-adapter and enables the model to more effectively bridge the multimodal domain gap during the early stage of training.

C MORE DETAILS IN OUR FINETUNING STRATEGY.

To clarify the motivation for adopting a Gaussian MAE ViT backbone, we provide a visual comparison of the masking mechanisms in Figure 7. As illustrated, Vanilla MAE performs hard masking by directly discarding a subset of tokens. However, in the context of cross-modal fine-tuning, hard masking removes tokens and introduces structural discontinuities, which often exacerbate the already irregular distribution of 3D tokens. In contrast, our Gaussian MAE adopts a soft masking strategy: all tokens are preserved, and their representations are modulated using multiplicative Gaussian noise (denoted as $\mathcal{N}$ in the figure). Our Gaussian modulation preserves the complete token sequence and acts as an effective regularizer against domain-specific noise, facilitating our 3D-to-2D adaptation.

We additionally employ Grad-CAM to analyze the token-wise importance distribution across the ViT hierarchy. The heatmaps are computed by weighting the spatial feature maps with the averaged gradients of the prediction score with respect to the token embeddings, highlighting the tokens most

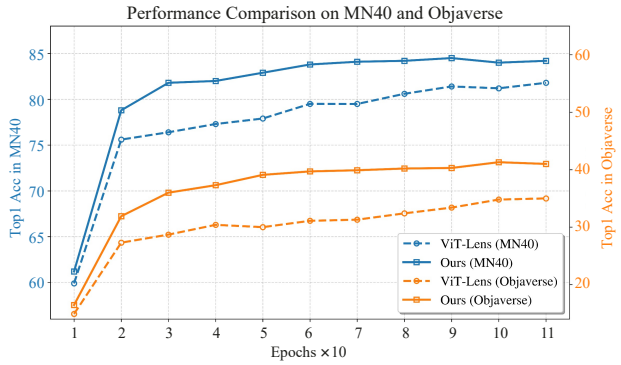


Figure 6: Comparison of the top 1 accuracy between our method and the ViT-Lens method under different training epochs.

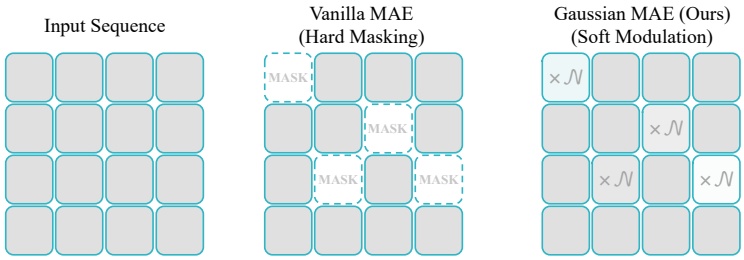


Figure 7: Comparison of masking mechanisms. In contrast to the binary hard masking used in Vanilla MAE (middle), our Gaussian MAE (right) applies soft modulation via multiplicative noise.

relevant to the model's decision. For visualization, we reshape the 512 tokens of each layer into 32×16 grids. As illustrated in Figure 8, the activation maps in the deeper layers (layers > 6) remain consistent, implying a stable semantic understanding. Conversely, shallow layers—particularly the first—exhibit marked variations in distribution. This discrepancy indicates that domain shift is predominantly addressed in the initial stages of the network, confirming that fine-tuning the shallow layers is reasonable.

D T-SNE Analysis of the Feature Transform.

To investigate the domain adaptation process within the 3D branch, we employ t-SNE to visualize feature distributions across key processing stages, as illustrated in Figure 9. We extract features after the point tokenizer, pre-adapter, post-adapter, and the shallow Transformer layer. Significant distributional shifts observed across the initial stages evidence the progressive transformation of raw 3D features. Notably, the distributions of the post-adapter and shallow ViT layers exhibit a high degree of alignment. This tendency indicates that our module effectively bridges the modality gap via projecting the original 3D group features into a representation space that is compatible with the pre-trained ViT.

E Additional Comparative Results on VQA.

Figure 10 presents a qualitative comparison with ViT-Lens on additional VQA tasks. Our method yields significantly more precise 3D scene descriptions and demonstrates better robustness to multi-modal inputs. In contrast, ViT-Lens is prone to hallucinations, frequently referencing non-existent objects or generating contextually inaccurate global descriptions. These results indicate that Ada3D could extract fine-grained features that align effectively with the text embedding space, facilitating seamless integration into LLMs for complex multimodal reasoning.

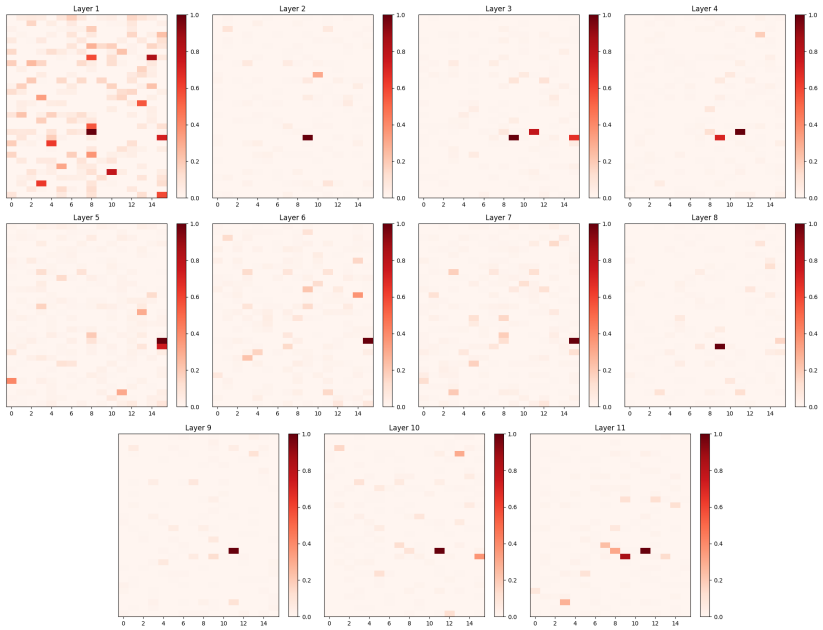
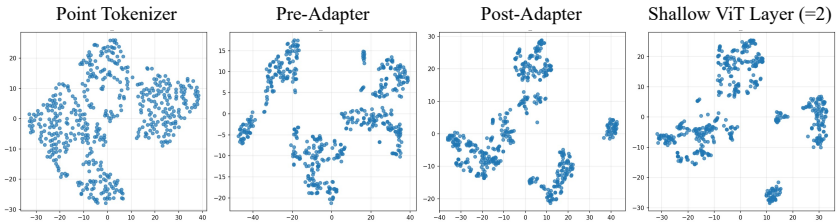
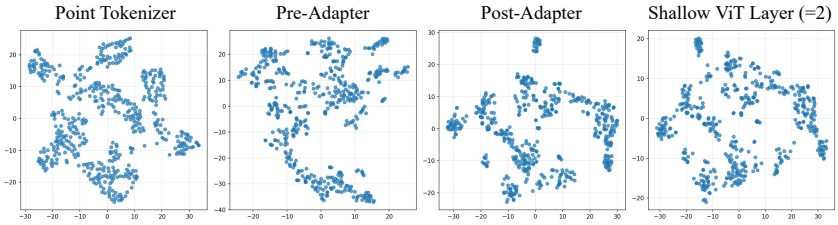


Figure 8: Grad-CAM heatmaps across ViT layers. Deep layers exhibit consistent attention patterns, while shallow layers show significant variation.



(a) T-SNE for object_0813



(b) T-SNE for object_0816

Figure 9: T-SNE visualization of feature adaptation. The clear alignment between post-adapter outputs and shallow ViT layers demonstrates that our method effectively bridges the modality gap.

F Impact of Projection View Numbers.

We further investigate the effect of the number of projection views V used in the Multi-View Aggregation (MVA) module. As presented in Table 8, we evaluate the zero-shot classification performance with 1, 3, and 6 projection views on three benchmarks: Objaverse-LVIS, ModelNet40, and ScanObjectNN. The results reveal that the MVA module exhibits an optimal configuration at 6 views, achieving the best performance across all three benchmarks. Compared to the single-view baseline, the 6-view aggregation provides improvements of $+2.1\%$ on Objaverse-LVIS, $+0.6\%$ on ModelNet40, and $+1.3\%$ on ScanObjectNN.



Figure 10: VQA comparison with ViT-Lens. Our method provides precise scene descriptions and demonstrates stronger robustness against multimodal hallucinations.

Table 8: Zero-shot Top-1 classification accuracy (%) under different numbers of projection views V in the MVA module.

V	Objaverse-LVIS	ModelNet40	ScanObjectNN
1	51.9	87.9	64.6
3	52.7	87.7	65.1
6	**54.0**	**88.5**	**65.9**