AVA-3D: ATTRIBUTE-BASED VISUAL ARBITRATION FOR UNIFIED 3D REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal pre-training for 3D point clouds has shown strong potential by aligning 3D shapes with corresponding images and language descriptions. However, existing methods often integrate multi-view visual information at the global or view level, overlooking heterogeneous visual attributes across viewpoints and their varying relevance to object geometry. To address this issue, we propose AVA-3D, an Attribute-based Visual Arbitration framework. AVA-3D introduces a Mixture of Arbitration (MoA) module to reorganize view-indexed features into complementary latent visual attributes through dual-axis competitive arbitration, together with a Conditional Attribute Fusion (CAF) module that adaptively routes and transforms these attributes according to 3D geometry. By enabling attribute-level organization across views and geometry-conditioned fusion across modalities, AVA-3D learns more discriminative 3D representations. Extensive experiments on Objaverse, ModelNet40, and ScanObjectNN demonstrate consistent improvements across downstream tasks.

1 INTRODUCTION

Multimodal 3D representation learning aims to align 3D data with visual and textual modalities in a unified embedding space, providing a foundation for emerging applications such as spatial computing (Chen et al., 2024), embodied intelligence (Zhen et al., 2024), and immersive environments (Hu et al., 2025).

Recent progress has largely been driven by transferring semantic knowledge from large-scale 2D vision-language models to the 3D domain. Projection-based methods (Zhang et al., 2022; Zhu et al., 2023) render point clouds into multi-view images to leverage pretrained 2D encoders, while contrastive methods (Xue et al., 2023; 2024; Liu et al., 2023; Zhou et al., 2024; Wang et al., 2024; Lobo et al., 2026) align images, text, and point clouds in a shared embedding space. Building on such cross-modal alignment, recent approaches further incorporate visual information into 3D representations, either through global visual cues (Gao et al., 2024; Zhou et al., 2025) or finer-grained view-level features (Lee et al., 2025) to provide complementary semantics.

Despite these advances, the integration of multi-view visual information into 3D representations is still constrained by two key challenges. First, different views provide partial observations and often reveal distinct visual attributes of the same object. Aggregating them into a unified representation may not fully preserve the diversity of visual evidence across viewpoints. Second, images emphasize appearance semantics, whereas point clouds capture geometric structure. The relevance of visual information therefore varies with the underlying 3D representation, making uniform cross-modal fusion less effective in exploiting their complementary cues. Together, such *view heterogeneity* and the *modality gap* hinder the effective integration of multi-view visual information into 3D representations. Although prior studies have investigated unified multi-view representations (Wang et al., 2023b), camera–LiDAR fusion (Wang et al., 2023a; Li et al., 2024), and prototype-based joint image–point cloud pre-training for scene-level perception (Chen et al., 2026), how to organize diverse visual attributes from partial views and adapt their contributions to object geometry remains underexplored in object-level multimodal 3D representation learning.

These observations motivate shifting visual enhancement from holistic view-level integration to a more fine-grained attribute-level formulation. Across views, heterogeneous observations should be

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.

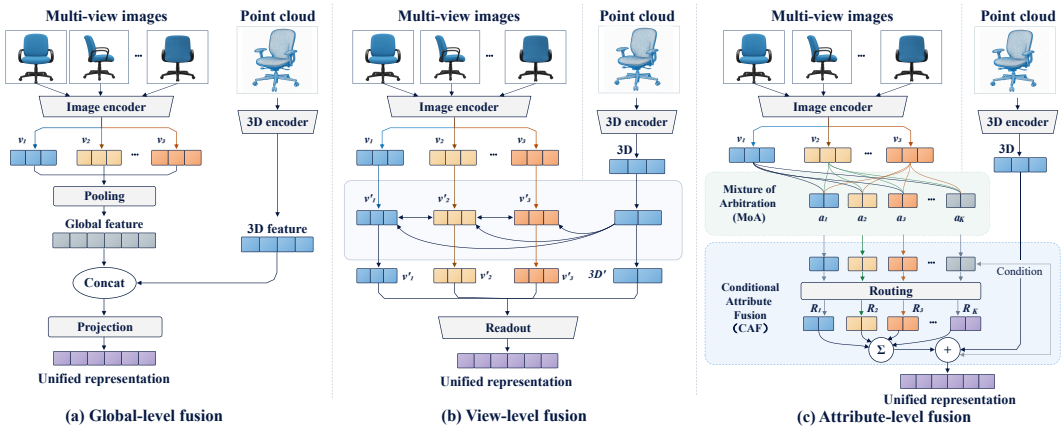

Figure 1: Comparison of visual fusion paradigms for unified 3D representation learning. (a) Global-level fusion holistically aggregates multi-view features. (b) View-level fusion models interactions across individual views. (c) Attribute-level fusion organizes multi-view observations into latent visual attributes and performs geometry-conditioned fusion.

organized into latent visual attributes, consolidating shared visual evidence while preserving complementary characteristics revealed by individual viewpoints. Across modalities, the contribution of each visual attribute should depend on the 3D representation, allowing geometry-relevant visual cues to be selectively emphasized during fusion.

To this end, we propose **AVA-3D**, an **A**ttribute-based **V**isual **A**rbitration framework for multimodal 3D representation learning. As illustrated in Fig. 1, AVA-3D first employs a Mixture of Arbitration (MoA) module that uses learnable slots to reorganize view-indexed features into attribute-indexed representations, thereby discovering latent visual attributes from complementary partial observations. The Conditional Attribute Fusion (CAF) module then conditions on the 3D representation to adaptively route these attributes and transform them into attribute-specific residual updates. By organizing heterogeneous multi-view information into visual attributes and adapting their contributions to object geometry, AVA-3D learns more discriminative 3D representations. Extensive experiments on Objaverse, ModelNet40, and ScanObjectNN demonstrate that AVA-3D consistently outperforms existing methods across diverse evaluation settings.

Our contributions are summarized as follows:

- We propose AVA-3D, an attribute-level visual arbitration framework that mitigates view heterogeneity and the modality gap in multimodal 3D representation learning.
- We introduce MoA to extract complementary latent visual attributes and CAF to adaptively integrate them with 3D representations for geometry-aware visual enhancement.
- Extensive experiments demonstrate state-of-the-art performance across multiple downstream benchmarks.

2 Related Work

2.1 Multimodal 3D Representation Learning

Projection-based methods convert point clouds into 2D representations to exploit pretrained image or vision-language models. PointCLIP (Zhang et al., 2022) projects point clouds into multi-view depth maps for CLIP-based recognition, while PointCLIP V2 (Zhu et al., 2023) improves compatibility with realistic depth projection and 3D-specific prompts. P2P (Wang et al., 2022) combines geometry-preserving projection with geometry-aware coloring, and OpenDlign (Mao et al., 2024) further narrows the domain gap through contour-aware depth alignment.

Contrastive methods align 3D representations with pretrained image-text spaces. CG3D (Hegde et al., 2023) learns from point cloud–image–text triplets with contrastive supervision and visual

prompt tuning. ULIP (Xue et al., 2023) aligns point-cloud features with a pretrained image-text space, while ULIP-2 (Xue et al., 2024) scales this formulation with generated descriptions. OpenShape (Liu et al., 2023) and Uni3D (Zhou et al., 2024) scale data and model capacity, whereas MRD (Wang et al., 2024) distills intra-modal and cross-modal relations. 3D-MRL (Lobo et al., 2026) further learns nested multimodal representations across multiple embedding dimensions. Beyond point clouds, CLIP-GS (Liao et al., 2025) extends alignment to 3D Gaussians, while GaussFusion (You et al., 2026) combines masked Gaussian modeling with image-text supervision.

Multi-view visual representations exploit complementary image observations for object-level 3D learning. DuoDuoCLIP (Lee et al., 2025) introduces cross-view attention into a CLIP-based encoder, MixCon3D (Gao et al., 2024) combines aggregated multi-view and point-cloud features, and OpenView (Zhou et al., 2025) preserves view-level embeddings with a sequence-independent fusion encoder. KR-MAE (Jin et al., 2026) selects reliable views and aligns salient 3D parts with image patches to reduce view redundancy.

Existing methods mainly encode multi-view information at the view or holistic level, leaving the organization of heterogeneous observations into complementary visual attributes underexplored.

2.2 Multi-view and Cross-modal Fusion

Multi-view fusion integrates observations through viewpoint selection, feature aggregation, and cross-view interaction. MVTN (Hamdi et al., 2021) learns task-dependent viewpoints, while MetaViewer (Wang et al., 2023b) learns a unified representation through view-specific reconstruction. For scene-level perception, UniFusion (Qin et al., 2023) performs spatial-temporal fusion in bird's-eye-view space, whereas MvACon (Liu et al., 2024) contextualizes image features across cameras for 2D-to-3D lifting.

Cross-modal fusion combines visual and geometric information for 3D perception. PointPainting (Vora et al., 2020) associates image semantics with LiDAR points, while BEVFusion (Liang et al., 2022) integrates camera and LiDAR features in bird's-eye-view space. UniTR (Wang et al., 2023a) employs a shared transformer for multimodal interaction, GAFusion (Li et al., 2024) introduces LiDAR-guided adaptive fusion, and IS-Fusion (Yin et al., 2024) combines instance-level information with scene context. CLAP (Chen et al., 2026) jointly pretrains image and point-cloud encoders with shared prototypes, while DA-FSS (Bian & Xu, 2026) decouples semantic and geometric experts for few-shot segmentation.

Existing cross-modal fusion methods mainly organize interactions at the point, region, or instance level, while geometry-conditioned integration of view-derived visual attributes remains underexplored in object-level multimodal 3D representation learning.

3 Method

3.1 Preliminaries

For the i-th 3D object, let $\mathcal{P}_i$, $\mathcal{T}_i$, and $\{\mathcal{I}_i^m\}_{m=1}^M$ denote its point cloud, text description, and M rendered views, respectively. Multimodal 3D representation learning maps these modalities into a shared embedding space to support downstream cross-modal tasks.

Following prior tri-modal contrastive frameworks Xue et al. (2023); Liu et al. (2023); Zhou et al. (2024); Hegde et al. (2023), let E_I and E_T denote frozen CLIP image and text encoders that extract pretrained visual and textual embeddings, and let E_P denote a trainable point-cloud encoder that maps point clouds into the shared embedding space. After modality-specific projection and ℓ_2 normalization, the resulting point-cloud, text, and view-wise image embeddings are denoted by $\mathbf{p}_i, \mathbf{t}_i, \mathbf{v}_i^m \in \mathbb{R}^d$. The image embeddings are stacked as $\mathbf{V}_i = [\mathbf{v}_i^1, \dots, \mathbf{v}_i^M]^\top \in \mathbb{R}^{M\times d}$.

Given two batches of normalized embeddings, $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^B$ and $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^B$, where embeddings with the same index correspond to the same object, the directional contrastive loss is defined as

$$\mathcal{L}_{\mathbf{X}\to\mathbf{Y}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(\mathbf{x}_i^\top\mathbf{y}_i/\tau)}{\sum_{j=1}^{B}\exp(\mathbf{x}_i^\top\mathbf{y}_j/\tau)}, \tag{1}$$

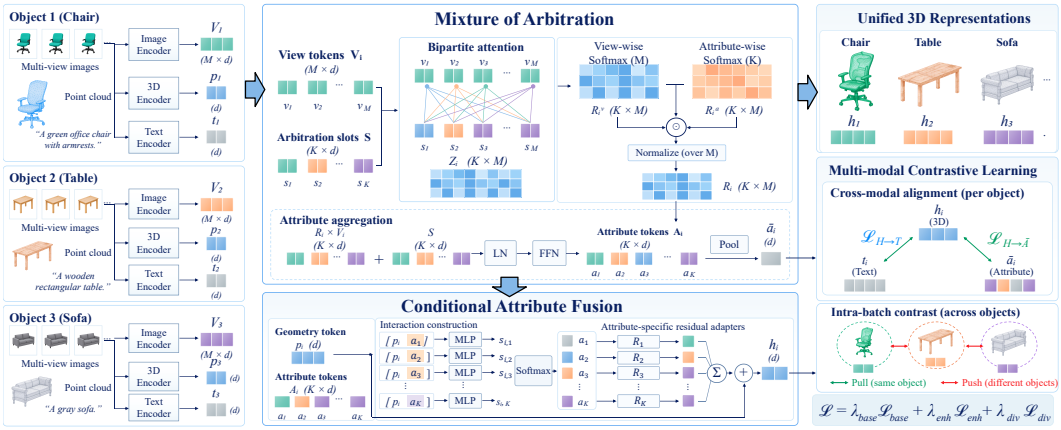

Figure 2: Overview of AVA-3D. MoA reorganizes multi-view features into latent visual attributes through dual-axis arbitration, while CAF conditionally routes and transforms these attributes to enhance the 3D representation. The model is trained with multimodal contrastive supervision.

where B denotes the batch size and $\tau > 0$ is a learnable temperature parameter. The corresponding symmetric contrastive loss is

$$\mathcal{L}_{\mathbf{X}\leftrightarrow\mathbf{Y}} = \frac{1}{2}\left(\mathcal{L}_{\mathbf{X}\to\mathbf{Y}} + \mathcal{L}_{\mathbf{Y}\to\mathbf{X}}\right). \tag{2}$$

In frameworks that combine multi-view images with point clouds, interactions between $\mathbf{V}_i$ and $\mathbf{p}_i$ enable visual information from partial observations to complement the geometric representation.

3.2 Mixture of Arbitration

Rendered views provide partial observations of the same object and expose diverse visual characteristics from different viewpoints. Direct aggregation may entangle such heterogeneous evidence, while single-axis attention can identify relevant views for each query but does not explicitly regulate how visual evidence is distributed across different latent attributes. To obtain a finer-grained visual representation, the Mixture of Arbitration (MoA) module reorganizes view-indexed observations into attribute-indexed representations through dual-axis competitive arbitration.

Specifically, MoA introduces K learnable arbitration slots $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_K]^\top \in \mathbb{R}^{K\times d}$ as latent visual attribute anchors. Given the multi-view representation $\mathbf{V}_i \in \mathbb{R}^{M\times d}$, the slot–view affinity is computed as

$$\mathbf{Z}_i = \frac{(\mathbf{S}\mathbf{W}_Q^a)(\mathbf{V}_i\mathbf{W}_K^a)^\top}{\sqrt{d}} \in \mathbb{R}^{K\times M}. \tag{3}$$

Instead of normalizing $\mathbf{Z}_i$ along a single dimension, MoA performs arbitration along both the view and attribute dimensions:

$$\begin{aligned} \mathbf{R}_i^v &= \mathrm{Softmax}_M(\mathbf{Z}_i), \\ \mathbf{R}_i^a &= \mathrm{Softmax}_K(\mathbf{Z}_i). \end{aligned} \tag{4}$$

The view-wise distribution $\mathbf{R}_i^v$ determines which observations provide relevant evidence for each latent attribute, whereas the attribute-wise distribution $\mathbf{R}_i^a$ determines how the evidence of each view is allocated among competing attributes. Their combination therefore favors attribute–view associations that are simultaneously relevant from the attribute perspective and distinctive from the view perspective.

The two arbitration distributions are combined by element-wise agreement and renormalized over the view dimension:

$$\widehat{\mathbf{R}}_i = \mathbf{R}_i^v \odot \mathbf{R}_i^a, \qquad [\mathbf{R}_i]_{k,m} = \frac{[\widehat{\mathbf{R}}_i]_{k,m}}{\sum_{n=1}^{M}[\widehat{\mathbf{R}}_i]_{k,n}}. \tag{5}$$

The resulting arbitration matrix $\mathbf{R}_i$ selectively routes multi-view evidence to different latent attributes. The routed evidence is then integrated with the corresponding arbitration slots to construct the attribute representation:

$$\begin{aligned}\widetilde{\mathbf{A}}_i &= \mathrm{LN}\left(\mathbf{S} + \mathbf{R}_i \mathbf{V}_i \mathbf{W}_V^a\right), \\ \mathbf{A}_i &= \mathrm{LN}\left(\widetilde{\mathbf{A}}_i + \mathrm{FFN}(\widetilde{\mathbf{A}}_i)\right),\end{aligned} \tag{6}$$

where $\mathbf{A}_i = [\mathbf{a}_{i,1}, \ldots, \mathbf{a}_{i,K}]^\top \in \mathbb{R}^{K\times d}$ contains K latent visual attributes. These attributes are not predefined semantic categories, but are learned end-to-end to capture complementary visual characteristics across objects.

For global visual alignment, the attribute tokens are summarized into a normalized embedding

$$\bar{\mathbf{a}}_i = \mathrm{Norm}\left(\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{a}_{i,k}\right)\mathbf{W}_O^a\right), \tag{7}$$

where $\mathrm{Norm}(\cdot)$ denotes ℓ_2 normalization. The complete attribute set $\mathbf{A}_i$ is retained for subsequent geometry-conditioned fusion.

3.3 Conditional Attribute Fusion

The complementary value of visual attributes varies with the underlying 3D geometry. For a given object, some visual cues may provide discriminative information beyond the point-cloud representation, whereas others may be redundant or weakly relevant. Uniformly fusing all attributes can therefore dilute useful cross-modal evidence. To enable finer-grained visual–3D interaction, the Conditional Attribute Fusion (CAF) module adaptively determines both the contribution of each visual attribute and how it modifies the 3D representation.

Specifically, for each visual attribute $\mathbf{a}_{i,k}$, its relation to the point-cloud representation is characterized by

$$\mathbf{x}_{i,k} = [\mathbf{p}_i;\ \mathbf{a}_{i,k};\ \mathbf{p}_i \odot \mathbf{a}_{i,k};\ |\mathbf{p}_i - \mathbf{a}_{i,k}|] \in \mathbb{R}^{4d}. \tag{8}$$

This interaction descriptor preserves modality-specific information while explicitly encoding agreement and discrepancy between the visual attribute and 3D geometry. A shared conditional router maps each interaction to an attribute score:

$$s_{i,k} = \mathbf{w}_R^\top \sigma\left(\mathrm{LN}(\mathbf{x}_{i,k})\mathbf{W}_R\right) + b_R, \qquad k = 1, \ldots, K, \tag{9}$$

where $\sigma(\cdot)$ denotes GELU. The routing weights are normalized over all visual attributes:

$$\boldsymbol{\alpha}_i = \mathrm{Softmax}\left([s_{i,1}, \ldots, s_{i,K}]\right) \in \mathbb{R}^K. \tag{10}$$

The coefficient $\alpha_{i,k}$ measures the relative contribution of the k-th visual attribute conditioned on the current 3D representation, allowing CAF to emphasize attributes that provide more complementary information to the encoded geometry.

Beyond determining which attributes should contribute, CAF further models how different attributes modify the 3D representation. Each latent attribute is associated with a lightweight low-rank residual adapter:

$$\Delta\mathbf{p}_{i,k} = \frac{\gamma_k}{r}\sigma\left(\mathrm{LN}(\mathbf{x}_{i,k})\mathbf{W}_{D,k}\right)\mathbf{W}_{U,k}, \tag{11}$$

where $\mathbf{W}_{D,k} \in \mathbb{R}^{4d\times r}$ and $\mathbf{W}_{U,k} \in \mathbb{R}^{r\times d}$ are the down- and up-projection matrices, $r \ll d$ is the bottleneck dimension, and γ_k is a learnable scaling factor. The up-projection matrix $\mathbf{W}_{U,k}$ is initialized to zero, allowing the residual branch to progressively learn attribute-specific transformations without perturbing the initial 3D representation.

The enhanced 3D representation is obtained by aggregating the routed residual updates:

$$\mathbf{h}_i = \mathrm{Norm}\left(\mathbf{p}_i + \sum_{k=1}^{K}\alpha_{i,k}\Delta\mathbf{p}_{i,k}\right), \tag{12}$$

where $\mathrm{Norm}(\cdot)$ denotes ℓ_2 normalization. In this way, CAF decouples attribute routing from attribute transformation: the routing coefficients determine the relative contribution of visual attributes, while the attribute-specific adapters determine how each attribute updates the 3D representation.

3.4 Overall Framework

By integrating MoA and CAF, AVA-3D forms a progressive visual enhancement framework, as illustrated in Fig. 2. For the i-th object, the point-cloud encoder produces the geometry-centered representation $\mathbf{p}_i$, while the frozen image encoder extracts the multi-view representation $\mathbf{V}_i$. MoA reorganizes $\mathbf{V}_i$ into latent visual attributes $\mathbf{A}_i$ through dual-axis competitive arbitration and summarizes them as $\bar{\mathbf{a}}_i$ for global visual alignment. CAF then conditions these attributes on $\mathbf{p}_i$, adaptively routing their contributions and transforming them into residual updates to obtain the enhanced representation $\mathbf{h}_i$. The frozen text encoder provides the textual embedding $\mathbf{t}_i$ for cross-modal alignment.

Contrastive supervision is applied to both the original and enhanced 3D representations. Let $\mathbf{P} = \{\mathbf{p}_i\}_{i=1}^{B}$, $\mathbf{H} = \{\mathbf{h}_i\}_{i=1}^{B}$, $\mathbf{T} = \{\mathbf{t}_i\}_{i=1}^{B}$, and $\bar{\mathbf{A}} = \{\bar{\mathbf{a}}_i\}_{i=1}^{B}$. The corresponding objectives are

$$\begin{aligned} \mathcal{L}_{\text{base}} &= \mathcal{L}_{\mathbf{P}\leftrightarrow\mathbf{T}} + \mathcal{L}_{\mathbf{P}\leftrightarrow\bar{\mathbf{A}}}, \\ \mathcal{L}_{\text{enh}} &= \mathcal{L}_{\mathbf{H}\leftrightarrow\mathbf{T}} + \mathcal{L}_{\mathbf{H}\leftrightarrow\bar{\mathbf{A}}}. \end{aligned} \tag{13}$$

A diversity regularizer $\mathcal{L}_{\text{div}}$ is further applied to the pre-softmax slot–view affinities in MoA to discourage correlated view-scoring patterns across arbitration slots. The overall training objective is

$$\mathcal{L} = \lambda_{\text{base}}\mathcal{L}_{\text{base}} + \lambda_{\text{enh}}\mathcal{L}_{\text{enh}} + \lambda_{\text{div}}\mathcal{L}_{\text{div}}, \tag{14}$$

where λ_{base}, λ_{enh}, and λ_{div} balance the corresponding objectives.

During training, the image and text encoders remain frozen, while the point-cloud encoder, projection heads, MoA, and CAF are jointly optimized. During inference, $\mathbf{h}_i$ serves as the unified 3D representation for downstream tasks.

4 Experiment

4.1 Experimental Setup

We follow the training data and evaluation protocols of prior works (Liu et al., 2023; Wang et al., 2024; Gao et al., 2024; Zhou et al., 2025). Three pre-training settings are considered: *Ensembled*, containing approximately 876K objects from Objaverse (Deitke et al., 2023), ShapeNet (Chang et al., 2015), 3D-FUTURE (Fu et al., 2021), and ABO (Collins et al., 2022); *Ensembled (No LVIS)*, which excludes objects overlapping with the Objaverse-LVIS test set; and *ShapeNet-only*, containing 52.5K objects from 55 categories. We directly use the multimodal data released by OpenShape (Liu et al., 2023), where each object is associated with 12 rendered views and textual descriptions generated by BLIP (Li et al., 2022) and refined by GPT-4 (OpenAI, 2024). No additional training data or preprocessing is introduced.

Zero-shot 3D classification is evaluated on Objaverse-LVIS (Deitke et al., 2023), containing 46,832 objects from 1,156 LVIS (Gupta et al., 2019) categories; ModelNet40 (Wu et al., 2015), containing 2,468 objects from 40 categories; and ScanObjectNN (Uy et al., 2019), containing 2,902 real-world objects from 15 categories. Top-1, Top-3, and Top-5 classification accuracies are reported.

AVA-3D is trained for 200 epochs on eight NVIDIA A800 GPUs using AdamW (Loshchilov, 2019), with an initial learning rate of 0.006, a 10-epoch warmup, cosine decay to 0.003, a batch size of 2048, and an EMA factor of 0.9994. PointBERT (Yu et al., 2022) serves as the 3D encoder, while OpenCLIP ViT-bigG-14 (Cherti et al., 2023) provides frozen image and text encoders. Unless otherwise specified, we use $K = 8$ arbitration slots and $M = 4$ input views. Additional implementation details are provided in the supplementary material.

4.2 Main Results

We compare AVA-3D with representative multimodal 3D representation learning methods under three pre-training settings, with results summarized in Table 1. Under the full *Ensembled* setting, AVA-3D achieves 55.8%, 87.7%, and 65.2% Top-1 accuracy on Objaverse-LVIS, ModelNet40, and ScanObjectNN, respectively. Compared with DuoDuoCLIP, which employs a substantially larger backbone, AVA-3D improves Top-1 accuracy by 1.7 percentage points on Objaverse-LVIS and 4.5 points on ScanObjectNN. More notably,using the same backbone as MixCon3D, AVA-3D yields

Table 1: Zero-shot 3D classification accuracy (%) under different pre-training settings. Best and second-best results within each setting are highlighted in **bold** and underline, respectively.

Method	Params (M)	Pre-training data	Objaverse-LVIS (Deitke et al., 2023)			ModelNet40 (Wu et al., 2015)			ScanObjectNN (Uy et al., 2019)		
			Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
PointCLIP (Zhang et al., 2022)	–	None	1.9	4.1	5.8	19.3	28.6	34.8	10.5	20.8	30.6
PointCLIP v2 (Zhu et al., 2023)	–	None	**4.7**	**9.5**	**12.9**	**63.6**	**77.9**	**85.0**	**42.2**	**63.3**	**74.5**
ReCon (Qi et al., 2023)	43.6	ShapeNet	1.1	2.7	3.7	61.2	73.9	78.1	42.3	62.5	75.6
CG3D (Hegde et al., 2023)	22.5	ShapeNet	5.0	9.5	11.6	48.7	60.7	66.5	42.5	57.3	60.8
CLIP2Point (Huang et al., 2023)	–	ShapeNet	2.7	5.8	7.9	49.5	71.3	81.2	25.5	44.6	59.4
ULIP (Xue et al., 2023)	22.6	ShapeNet	6.2	13.6	17.9	60.4	79.0	84.4	51.5	71.1	80.2
OpenShape (Liu et al., 2023)	32.3	ShapeNet	10.8	20.2	25.0	70.3	86.9	91.3	51.3	69.4	78.4
TAMM (Zhang et al., 2024)	–	ShapeNet	13.7	24.2	29.2	73.1	88.5	91.9	**54.8**	74.5	83.3
MixCon3D (Gao et al., 2024)	32.5	ShapeNet	22.3	37.5	44.3	72.6	87.1	91.3	52.6	69.9	78.7
AVA-3D (Ours)	32.5	ShapeNet	**34.9**	**53.4**	**61.0**	**78.6**	**94.0**	**97.0**	54.0	**75.1**	**84.5**
ULIP (Xue et al., 2023)	22.6	Ensembled (no LVIS)	21.4	38.1	46.0	71.4	84.4	89.2	46.0	66.1	76.4
OpenShape (Liu et al., 2023)	32.3	Ensembled (no LVIS)	39.1	60.8	68.9	85.3	96.2	97.4	47.2	72.4	84.7
TAMM (Zhang et al., 2024)	–	Ensembled (no LVIS)	42.0	63.6	71.7	86.3	96.6	98.1	56.7	78.3	86.1
Uni3D-L (Zhou et al., 2024)	306.7	Ensembled (no LVIS)	46.2	67.6	74.7	86.6	96.3	97.8	58.4	81.4	90.1
MixCon3D (Gao et al., 2024)	32.5	Ensembled (no LVIS)	47.5	69.0	76.2	**87.3**	96.8	98.1	57.7	80.7	89.8
ViT-Lens-G (Lei et al., 2024)	–	Ensembled (no LVIS)	50.1	71.3	78.1	86.8	96.8	97.8	59.8	79.3	87.7
DuoduoCLIP (Lee et al., 2025)	87.0	Ensembled (no LVIS)	51.3	73.1	79.9	–	–	–	–	–	–
AVA-3D (Ours)	32.5	Ensembled (no LVIS)	**53.4**	**74.4**	**80.9**	86.8	**97.6**	**99.1**	**65.7**	**86.0**	**92.5**
ULIP (Xue et al., 2023)	22.6	Ensembled	26.8	44.8	52.6	75.1	88.1	93.2	51.6	72.5	82.3
OpenShape (Liu et al., 2023)	32.3	Ensembled	46.8	69.1	77.0	84.4	96.5	98.0	52.2	79.7	88.7
TAMM (Zhang et al., 2024)	–	Ensembled	50.7	73.2	80.6	85.0	96.6	98.1	55.7	80.7	88.9
ViT-Lens-G (Lei et al., 2024)	–	Ensembled	52.0	73.3	79.9	87.6	96.6	98.4	60.1	81.0	90.3
MixCon3D (Gao et al., 2024)	32.5	Ensembled	52.5	74.5	81.2	86.8	96.9	98.3	58.6	80.3	89.2
Uni3D-L (Zhou et al., 2024)	306.7	Ensembled	53.1	75.0	81.5	86.3	96.8	98.3	58.2	81.8	89.4
MRD-L (Wang et al., 2024)	306.7	Ensembled	53.6	75.7	82.4	87.3	**97.4**	98.8	64.1	**85.7**	91.9
OpenView (Zhou et al., 2025)	–	Ensembled	52.7	74.6	81.1	87.5	97.0	98.5	58.5	73.5	78.6
DuoduoCLIP (Lee et al., 2025)	87.0	Ensembled	54.1	76.0	82.4	–	–	–	60.7	82.4	88.5
AVA-3D (Ours)	32.5	Ensembled	**55.8**	**77.2**	**83.3**	**87.7**	**97.4**	**99.0**	**65.2**	85.6	**92.6**

Table 2: Component ablation on Objaverse-LVIS with ShapeNet-only training. Accuracy (%); best results are in **bold**.

Variant	Top-1	Top-3	Top-5
Baseline	26.0	43.1	50.3
Baseline + CAF	28.1	45.6	52.8
Baseline + MoA	33.1	51.8	59.4
Full (MoA + CAF)	**34.9**	**53.4**	**61.0**

Table 3: Effect of the number of arbitration slots on Objaverse-LVIS with ShapeNet-only training. Accuracy (%); best results are in **bold**.

Slots	Top-1	Top-3	Top-5
1	28.1	45.6	52.8
2	30.0	48.0	55.5
4	32.4	50.9	58.4
8	**34.9**	**53.4**	**61.0**
12	34.7	52.6	59.6

gains of 3.3, 0.9, and 6.6 percentage points on the three benchmarks, respectively. These improvements demonstrate the benefit of organizing multi-view observations into visual attributes and adapting their contributions to the underlying 3D geometry.

AVA-3D also maintains strong performance when the available pre-training data are reduced. Under *Ensembled (No LVIS)*, it achieves the best results on Objaverse-LVIS and ScanObjectNN, while remaining competitive on ModelNet40. The advantage becomes particularly pronounced in the more restricted *ShapeNet-only* setting: AVA-3D reaches 34.9% Top-1 accuracy on Objaverse-LVIS, outperforming the strongest baseline by 12.6 percentage points, and improves ModelNet40 Top-1 accuracy to 78.6%. These results indicate that fine-grained visual arbitration and geometry-conditioned fusion remain effective even with substantially less pre-training data, suggesting better utilization of the available multimodal supervision.

4.3 Ablation Studies

Components. To evaluate the contributions of MoA and CAF, we conduct component-wise ablations under the ShapeNet-only setting and report Top-1, Top-3, and Top-5 zero-shot classification accuracy on Objaverse-LVIS, where the performance differences are most pronounced. MixCon3D is adopted as the baseline due to its explicit fusion of multi-view visual and point-cloud representations. The results are summarized in Table 2. Our reproduced baseline achieves 26.0% Top-1 accuracy, slightly higher than the officially reported 22.3%. Introducing MoA alone substantially improves Top-1 accuracy to 33.1%, demonstrating the benefit of reorganizing heterogeneous multi-

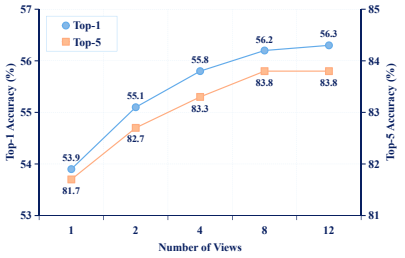

Figure 3: Effect of input-view count on zero-shot accuracy.

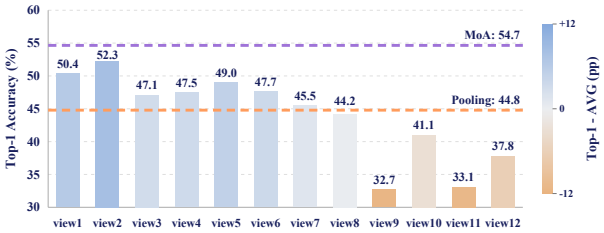

Figure 4: View-wise accuracy across viewpoints.

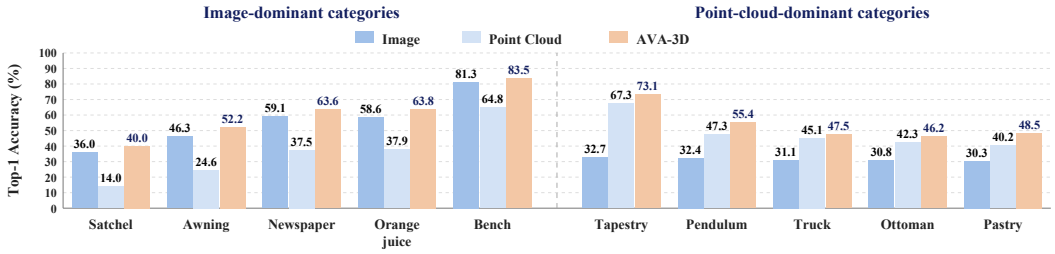

Figure 5: Category-wise comparison of image, point-cloud, and AVA-3D representations.

view observations into finer-grained visual attributes. Adding CAF alone yields a more moderate improvement to 28.1%, indicating that geometry-conditioned fusion can better exploit complementary visual information, while its effectiveness remains limited when the input visual representation still mixes heterogeneous view cues. Combining MoA and CAF achieves the best performance of 34.9% Top-1 accuracy, an 8.9-point improvement over the reproduced baseline, demonstrating that attribute-level visual arbitration and geometry-conditioned fusion provide complementary gains when jointly applied.

Number of arbitration slots. We further investigate the effect of the number of arbitration slots K under the ShapeNet-only setting, with results reported in Table 3. Performance consistently improves as K increases from 1 to 8, with Top-1 accuracy rising from 28.1% to 34.9% and Top-5 accuracy from 52.8% to 61.0%. A small number of slots provides insufficient capacity to organize the diverse visual characteristics exposed by multiple viewpoints. Increasing K allows MoA to represent these observations at a finer attribute granularity and yields steady performance gains. However, further increasing the number of slots to 12 brings no additional benefit, with Top-1 and Top-5 accuracy decreasing slightly to 34.7% and 59.6%, respectively. This suggests that excessive slots may fragment the available visual evidence or introduce redundant attribute representations. We therefore adopt $K = 8$ as the default setting.

Number of input views. To evaluate the generalization of AVA-3D to varying numbers of input views, we use the model pretrained with 4 views and vary the number of views only at inference time. As shown in Fig. 3, AVA-3D generalizes well across different view counts without retraining. Top-1 accuracy increases from 53.9% with a single view to 56.3% with 12 views, while Top-5 accuracy improves from 81.7% to 83.8% with eight views and then saturates. These results indicate that AVA-3D can effectively exploit additional visual observations beyond those seen during training. For fair comparison with prior methods using a limited number of views, such as MixCon3D (Gao et al., 2024) (4F) and DuoDuoCLIP (Lee et al., 2025) (5F), we adopt 4 input views as the default setting.

4.4 Discussion

View heterogeneity. To further investigate view heterogeneity and isolate the effect of multi-view aggregation, we remove CAF and evaluate the pretrained model using each OpenShape-rendered view individually. As shown in Fig. 4, Top-1 accuracy varies substantially across viewpoints, from

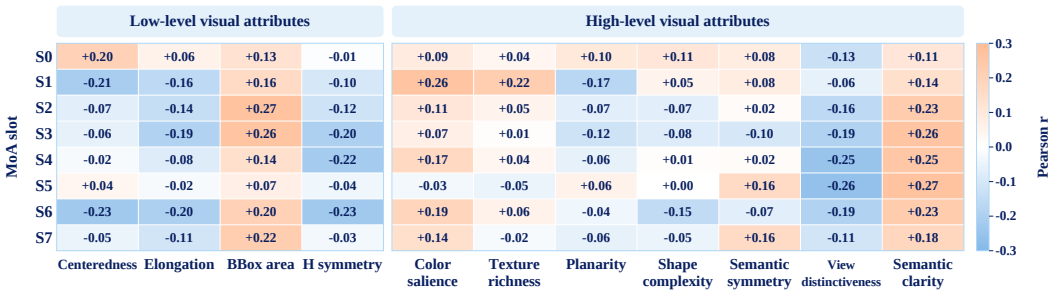

Figure 6: Correlations between MoA slot responses and low- and high-level visual attributes.

32.7% to 52.3%, indicating clear differences in the discriminative information provided by individual views. Aggregating all 12 views with pooling achieves 44.8%, which remains below the best single-view result. In contrast, applying MoA to the same 12 views improves Top-1 accuracy to 54.7%, outperforming both pooling-based aggregation and every individual view. These results demonstrate that MoA effectively organizes heterogeneous multi-view evidence and exploits complementary information across viewpoints.

Modality gap. To further examine the modality gap, we compare category-wise Top-1 accuracy using image features, point-cloud features, and the fused representation produced by CAF. As shown in Fig. 5, different object categories exhibit markedly different modality preferences. For example, *Orange juice* is substantially more discriminative in the image modality than in the point-cloud modality (58.6% vs. 37.9%), whereas *Tapestry* shows the opposite trend (32.7% vs. 67.3%). Despite these varying modality dependencies, CAF consistently achieves the strongest performance across the selected categories by adapting visual contributions to the underlying geometry. These results highlight the modality gap in object-level 3D recognition and demonstrate the effectiveness of geometry-conditioned fusion in exploiting complementary cross-modal information.

Attribute Analysis. To examine whether different MoA slots capture distinct visual attributes, we define a set of interpretable low- and high-level visual attributes and analyze their correlations with individual slot responses. As shown in Fig. 6, different slots exhibit distinct response patterns across these factors. For example, S0 responds more strongly to centeredness, while S2 and S3 show higher correlations with bounding-box area. In contrast, S1 is more responsive to high-level attributes such as color salience and texture richness, whereas S5 shows stronger responses to semantic symmetry and semantic clarity. Similar differences can be observed across the remaining slots. These results indicate that MoA encourages different slots to capture complementary visual characteristics rather than converging to homogeneous representations.

5 Conclusion

We presented AVA-3D, an attribute-based visual arbitration framework for multimodal 3D representation learning. Its Mixture of Arbitration (MoA) module organizes multi-view observations into complementary latent visual attributes through dual-axis competitive arbitration. The Conditional Attribute Fusion (CAF) module then routes and transforms these attributes conditioned on the 3D representation, adapting their contributions to object geometry. Extensive experiments across multiple pre-training settings and downstream benchmarks demonstrate consistent improvements, demonstrating their effectiveness.

Limitations. The visual attributes learned by AVA-3D remain latent, with no explicit constraints on their semantic meanings, leaving room for improved interpretability and controllability. The current framework also focuses primarily on object-level representation learning with rendered multi-view images. Future work will explore more interpretable attribute modeling and extend attribute arbitration to richer visual inputs and scene-level 3D understanding.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation of the manuscript. Additionally, we used these tools for language polishing to improve clarity and readability. The authors reviewed and revised all AI-assisted text to verify its accuracy and preserve the intended technical meaning. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Wentao Bian and Fenglei Xu. Rethinking multimodal few-shot 3d point cloud segmentation: From fused refinement to decoupled arbitration. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*, pp. 917–925, 2026.

Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *CoRR*, abs/1512.03012, 2015.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024.

Runjian Chen, Hang Zhang, Avinash Ravichandran, Wenqi Shao, Alex Wong, and Ping Luo. CLAP: unsupervised 3d representation learning for fusion 3d perception via curvature sampling and prototype learning. In *International Conference on Learning Representations*, 2026.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

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 *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 21094–21104, 2022.

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 *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 13142–13153, 2023.

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

Yipeng Gao, Zeyu Wang, Wei-Shi Zheng, Cihang Xie, and Yuyin Zhou. Sculpting holistic 3d representation in contrastive language-image-3d pre-training. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 22998–23008, 2024.

Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5351–5359, 2019.

Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. MVTN: multi-view transformation network for 3d shape recognition. In *IEEE International Conference on Computer Vision*, pp. 1–11, 2021.

Deepti Hegde, Jeya Maria Jose Valanarasu, and Vishal Patel. Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2028–2038, 2023.

Anjun Hu, Richard Tomsett, Valentin Gourmet, Massimo Camplani, Jas Kandola, and Hanting Xie. MiDSummer: Multi-guidance diffusion for controllable zero-shot immersive gaussian splatting scene generation. In *IEEE International Conference on Computer Vision*, pp. 26793–26805, 2025.

Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson WH Lau, Wanli Ouyang, and Wangmeng Zuo. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. In *IEEE International Conference on Computer Vision*, pp. 22100–22110, 2023.

Xianglong Jin, Zheng Wang, Rong Wang, and Feiping Nie. Reliable-view 2d-3d key-part aligned transformer with reinforced masking for 3d point cloud understanding. In *AAAI Conference on Artificial Intelligence*, pp. 5530–5538, 2026.

Han-Hung Lee, Yiming Zhang, and Angel Chang. Duoduo clip: Efficient 3d understanding with multi-view images. In *International Conference on Learning Representations*, volume 2025, pp. 48070–48091, 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 26637–26647, 2024.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900, 2022.

Xiaotian Li, Baojie Fan, Jiandong Tian, and Huijie Fan. Gafusion: Adaptive fusing lidar and camera with multiple guidance for 3d object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 21209–21218, 2024.

Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. BEVFusion: A simple and robust lidar-camera fusion framework. In *Conference and Workshop on Neural Information Processing Systems*, volume 35, pp. 10421–10434, 2022.

Guibiao Liao, Jiankun Li, Zhenyu Bao, Xiaoqing Ye, Qing Li, and Kanglin Liu. Clip-gs: Clip-informed gaussian splatting for view-consistent 3d indoor semantic understanding. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(8):1–24, 2025.

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. In *Conference and Workshop on Neural Information Processing Systems*, 2023.

Xianpeng Liu, Ce Zheng, Ming Qian, Nan Xue, Chen Chen, Zhebin Zhang, Chen Li, and Tianfu Wu. Multi-view attentive contextualization for multi-view 3d object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 16688–16698, 2024.

Márcus Lobo, Vitor Pereira Matias, Jeová Farias, and Moacir Ponti. 3d-mrl: Nested multimodal 3d representations via matryoshka representation learning. In *British Machine Vision Conference*, 2026.

I Loshchilov. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Ye Mao, Junpeng Jing, and Krystian Mikolajczyk. Opendlign: Open-world point cloud understanding with depth-aligned images. In *Conference and Workshop on Neural Information Processing Systems*, 2024.

OpenAI. Gpt-4 technical report, 2024. URL `https://arxiv.org/abs/2303.08774`.

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*, volume 202, pp. 28223–28243, 2023.

Zequn Qin, Jingyu Chen, Chao Chen, Xiaozhi Chen, and Xi Li. Unifusion: Unified multi-view fusion transformer for spatial-temporal representation in bird's-eye-view. In *IEEE International Conference on Computer Vision*, pp. 8656–8665, 2023.

Yue Ruan, Han-Hung Lee, Yiming Zhang, Ke Zhang, and Angel X. Chang. Tricolo: Trimodal contrastive loss for text to shape retrieval. pp. 5803–5813, 2024.

Chuan Tang, Xi Yang, Bojian Wu, Zhizhong Han, and Yi Chang. Parts2words: Learning joint embedding of point clouds and texts by bidirectional matching between parts and words. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6884–6893, 2023.

Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Duc Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *IEEE International Conference on Computer Vision*, pp. 1588–1597, 2019.

Sourabh Vora, Alex H. Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4603–4611, 2020.

Haiyang Wang, Hao Tang, Shaoshuai Shi, Aoxue Li, Zhenguo Li, Bernt Schiele, and Liwei Wang. Unitr: A unified and efficient multi-modal transformer for bird's-eye-view representation. In *IEEE International Conference on Computer Vision*, pp. 6769–6779, 2023a.

Huiqun Wang, Yiping Bao, Panwang Pan, Zeming Li, Xiao Liu, Ruijie Yang, and Di Huang. Multi-modal relation distillation for unified 3d representation learning. In *European Conference on Computer Vision*, 2024.

Ren Wang, Haoliang Sun, Yuling Ma, Xiaoming Xi, and Yilong Yin. Metaviewer: Towards a unified multi-view representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11590–11599, 2023b.

Ye Wang, Bowei Jiang, Changqing Zou, and Rui Ma. MXM-CLR: A unified framework for contrastive learning of multifold cross-modal representations. *CoRR*, abs/2303.10839, 2023c.

Ziyi Wang, Xumin Yu, Yongming Rao, Jie Zhou, and Jiwen Lu. P2P: tuning pre-trained image models for point cloud analysis with point-to-pixel prompting. In *Conference and Workshop on Neural Information Processing Systems*, 2022.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *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 *IEEE 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, and Silvio Savarese. ULIP-2: towards scalable multimodal pre-training for 3d understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 27081–27091, 2024.

Junbo Yin, Jianbing Shen, Runnan Chen, Wei Li, Ruigang Yang, Pascal Frossard, and Wenguan Wang. IS-Fusion: Instance-scene collaborative fusion for multimodal 3d object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 14905–14915, 2024.

Zhixuan You, Jihua Zhu, Yiding Sun, Zihao Guo, Haozhe Cheng, Dongxu Zhang, Lin Chen, and Hainan Luo. Gaussfusion: Towards multimodal 3d gaussian pretraining. *CoRR*, abs/2607.05906, 2026.

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 *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 19291–19300, 2022.

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 *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8542–8552, 2022.

Zhihao Zhang, Shengcao Cao, and Yu-Xiong Wang. TAMM: triadapter multi-modal learning for 3d shape understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 21413–21423, 2024.

Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. In *International Conference on Machine Learning*, pp. 61229–61245, 2024.

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*, 2024.

Ziyang Zhou, Pinghui Wang, Zi Liang, Haitao Bai, and Ruofei Zhang. Cross-modal 3d representation with multi-view images and point clouds. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3728–3739, 2025.

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 *IEEE International Conference on Computer Vision*, pp. 2639–2650, 2023.

A APPENDIX

A.1 MORE EXPERIMENTAL DETAILS

Training. A complete AVA-3D training run takes approximately 15.2 hours on eight NVIDIA A800-40GB GPUs, corresponding to 122 A800 GPU-hours. AVA-3D optimizes 32.5M trainable parameters, excluding the frozen OpenCLIP image and text encoders. For reference, Uni3D (Zhou et al., 2024) reports approximately 480 A100 GPU-hours with about 1B trainable parameters, while DuoDuoCLIP (Lee et al., 2025) requires 57 A5000 GPU-hours with 87M parameters. AVA-3D therefore uses substantially fewer trainable parameters, with $30.8\times$ fewer than Uni3D and $2.7\times$ fewer than DuoDuoCLIP, demonstrating favorable computational efficiency with a lightweight trainable architecture.

Attribute Response Analysis Settings. To characterize the visual factors captured by different MoA slots, we measure the correlation between slot–view responses and a set of interpretable view attributes, with results shown in Fig. 6. For each rendered view, four low-level visual attributes are computed directly from the image: *centeredness*, measuring the distance of the object from the image center; *elongation*, describing its spatial aspect ratio; *bounding-box area*, measuring relative foreground coverage; and *horizontal symmetry*, measuring left–right visual symmetry. We further consider seven high-level attributes, including *color salience*, *texture richness*, *planarity*, *shape complexity*, *semantic symmetry*, *view distinctiveness*, and *semantic clarity*. These descriptors characterize appearance, semantics, and view distinctiveness using OpenCLIP-based semantic scores and feature-space distances. For each object i, slot k, and attribute q, Pearson's r is computed across the 12 rendered views between the pre-softmax slot–view logits $\{[\mathbf{Z}_i]_{k,m}\}_{m=1}^{12}$ and the corresponding attribute scores $\{u_{i,m}^{q}\}_{m=1}^{12}$. The reported value is the mean correlation across objects. Detailed results and analysis are provided in Sec. 4.4.

Multimodal Data Preparation. For fair comparison, we directly use the OpenShape-prepared data adopted by prior works, without introducing additional data or preprocessing. OpenShape (Liu et al., 2023) constructs the multimodal training data from Objaverse, ShapeNetCore, 3D-FUTURE, and ABO. For each object, 10,000 points are sampled from the mesh surface, with RGB values interpolated from mesh textures, and 12 color images at a resolution of 512×512 are rendered from predefined viewpoints. Textual descriptions are collected from dataset metadata, including object names for Objaverse; shape, category, and subcategory names for ShapeNetCore; category, style, theme, and material information for 3D-FUTURE; and item names and product types for ABO. Uninformative metadata, such as filenames, identifiers, and timestamps, are filtered using GPT-4 while retaining the corresponding 3D objects. Additional captions are generated by BLIP and Microsoft Azure captioning from object thumbnails, when available, or fixed frontal renderings otherwise. To diversify textual supervision, visually similar images are further retrieved from LAION-5B using a CLIP ViT-L index, and their associated captions are added to the text candidates. Image and text embeddings are extracted with frozen OpenCLIP ViT-bigG-14 encoders and cached before training. The resulting data associate each object with a colored point cloud, multi-view visual embeddings, and a pool of filtered metadata descriptions, generated captions, and retrieved texts.

A.2 ADDITIONAL EXPERIMENTAL RESULTS

To further evaluate the unified 3D representations learned by AVA-3D, we provide additional cross-modal retrieval results and further analyses of the visual attributes captured by MoA.

Attribute Intervention Analysis. To further verify that different MoA slots respond to distinct visual attributes, we perform controlled interventions on the view preferred by each slot. For each slot, we select a representative attribute identified by the correlation analysis in Fig. 6 and modify the corresponding visual factor in its top-ranked view, such as shape complexity, color salience, bounding-box area, planarity, semantic symmetry, or semantic clarity. We then recompute the arbitration weights and examine whether the preferred view changes. As illustrated in Fig. 7, targeted attribute perturbations frequently trigger clear routing switches, with the original top-ranked view being replaced by another observation. For example, reducing the bounding-box area for S2 decreases the original top-ranked view from rank #1 to #12 and induces a reselection, with an output

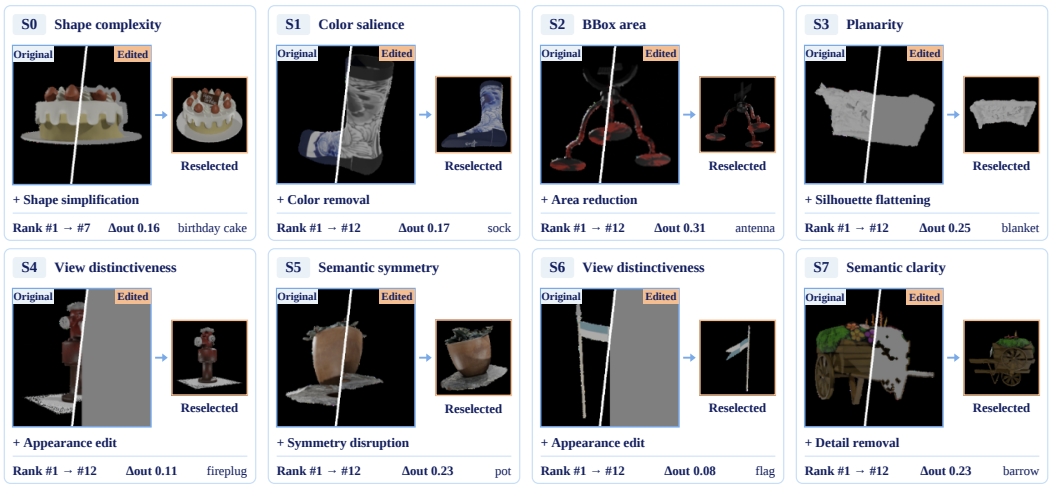

Figure 7: Controlled attribute interventions on MoA slots. Perturbing slot-correlated visual factors induces consistent routing shifts across views, providing evidence of differentiated attribute sensitivity among slots.



Figure 8: Qualitative image-to-shape retrieval results. Given real-world images as queries, AVA-3D retrieves semantically and visually consistent 3D models from Objaverse across diverse categories.

change of 0.31. Similarly, silhouette flattening for S3 and semantic-detail removal for S7 shift the perturbed views from rank #1 to #12, with output changes of 0.25 and 0.23, respectively. Consistent routing changes are also observed for shape simplification (S0), color removal (S1), appearance edits (S4 and S6), and symmetry disruption (S5). Notably, seven of the eight interventions demote the original preferred view to rank #12, while S0 drops from rank #1 to #7. These results demonstrate the association between MoA slots and differentiated visual factors, indicating that the learned slots actively leverage these attributes when arbitrating multi-view observations.

Qualitative Retrieval Analysis. Following prior works (Liu et al., 2023; Wang et al., 2024; Zhou et al., 2024), we perform image-to-shape retrieval using real-world images collected from Flickr as queries and retrieve the most relevant 3D shapes from Objaverse. As shown in Fig. 8, AVA-3D retrieves shapes that are visually and semantically consistent with the query images, demonstrating effective image–3D alignment. We further conduct text-to-shape retrieval, with qualitative results shown in Fig. 9. AVA-3D also retrieves shapes that closely match the corresponding textual descriptions, further validating its cross-modal generalization.

Table 4: Text-to-shape retrieval performance on Text2Shape. * denotes reproduced results.

Method	Data	RR@1	RR@5
Parts2Words (Tang et al., 2023)	T2S	12.72	32.98
TriCoLo (Ruan et al., 2024)	T2S	12.22	32.23
MXM-CLR (Wang et al., 2023c)	T2S	16.83	39.06
ZeroShot B-32	Images	11.90	27.85
FT B-32	Images	15.01	32.46
OpenShape† (Liu et al., 2023)	Ens.	10.53	25.94
Uni3D-G (Zhou et al., 2024)	Ens.	10.78	26.37
DuoDuoCLIP* (5F) (Lee et al., 2025)	Ens.	14.17	31.28
AVA-3D (Ours)	–	**15.58**	**33.43**



Figure 9: Qualitative text-to-shape retrieval results on Objaverse.

Quantitative Retrieval Analysis. Following DuoDuoCLIP (Lee et al., 2025), we further evaluate text-to-shape retrieval quantitatively, with results reported in Table 4. Although AVA-3D is trained with four input views, we evaluate it under the same five-view setting as DuoDuoCLIP and reproduce DuoDuoCLIP using its official implementation. AVA-3D consistently outperforms previous methods, surpassing the 1B-parameter Uni3D-G and the approximately $3\times$ larger DuoDuoCLIP by 4.8 and 1.41 points in RR@1, respectively. These results demonstrate that AVA-3D achieves stronger cross-modal alignment with a substantially more lightweight representation framework.