Chorus: Multi-Teacher Pretraining for Holistic 3D Gaussian Scene Encoding

Anonymous CVPR submission

Paper ID ****

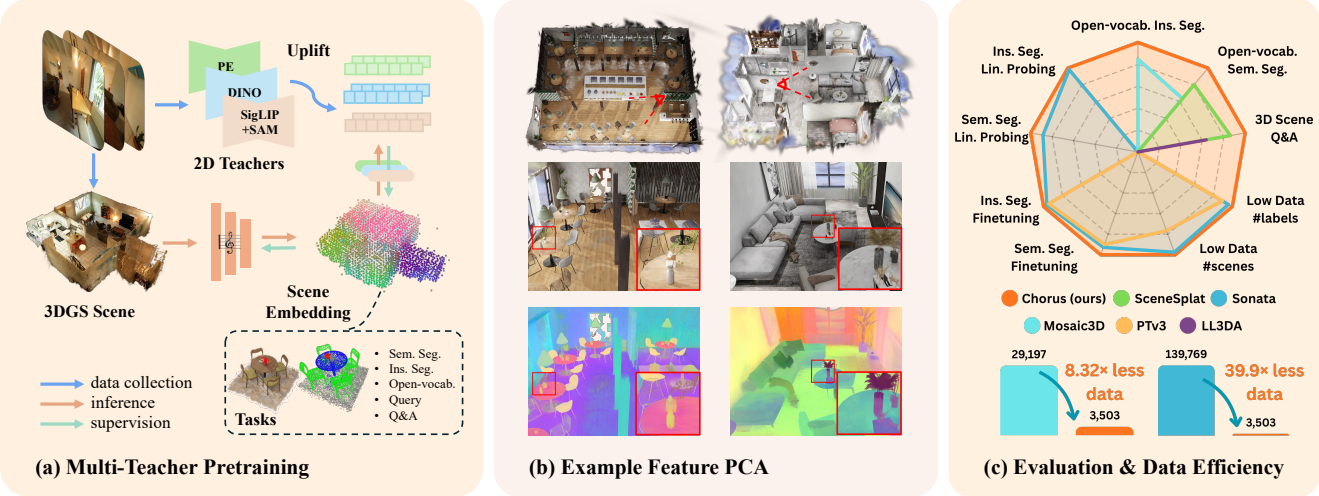

Figure 1. **Chorus Framework. (a) Multi-Teacher Pretraining.** A feed-forward 3DGS scene encoder with per-teacher projectors distills complementary signals—language-aligned, generalist, and object-aware—into a shared embedding. **(b) Example Feature PCA (results on novel scenes).** At inference we input the full 3DGS scene; PCA on encoder features presents clear semantic awareness despite domain shift. **(c) Evaluation & Data Efficiency.** Chorus attains strong results across scene understanding tasks while using noticeably fewer training scenes—8.32× and 39.9× less than the SoTA point clouds pretraining baselines—highlighting the efficiency of our pretraining.

Abstract

While 3DGS has emerged as a high-fidelity scene representation, encoding rich, general-purpose features directly from its primitives remains under-explored. We address this gap by introducing Chorus, a multi-teacher pretraining framework that learns a holistic feed-forward 3D Gaussian Splatting (3DGS) scene encoder by distilling complementary signals from 2D foundation models. Chorus employs a shared 3D encoder and teacher-specific projectors to learn from language-aligned, generalist, and object-aware teachers, encouraging a shared embedding space that captures signals from high-level semantics to fine-grained structure.

We evaluate Chorus on a wide range of tasks: open-vocabulary semantic and instance segmentation, linear and decoder probing, data-efficient supervision, as well as LLM-based Q&A. Besides 3DGS, we also test Chorus on several benchmarks that only support point clouds by pretraining a variant using only Gaussians' centers, colors, estimated normals. Interestingly, this encoder shows strong transfer and outperforms the point clouds baseline while using 39.9× fewer training scenes. Finally, we propose a render-and-distill adaptation that facilitates out-of-domain finetuning. Our code and model is released at this codebase.

1. Introduction

The community has made rapid progress on scene representations that enable photorealistic rendering, from neural radiance fields (NeRFs) [32] to real-time 3D Gaussian Splatting (3DGS) [22]. In parallel, there is a growing body of work that attaches semantic cues to these representations (*e.g.*, via attaching vision–language features [23, 31, 35, 39, 67]). Yet comparatively minor attention has been paid to treating the 3D representation itself as a modality from which we can directly mine general-purpose, transferable features at scale. 3DGS is particularly attractive in this regard: it preserves geometry–appearance primitives and supports fast differentiable rendering, which together make it a promising substrate for large-scale pretraining beyond view synthesis [6, 19, 28, 50].

We address the gap in generalizable 3DGS scene encoding by proposing Chorus–a multi-teacher pretraining

framework for training a native 3DGS encoder to align with complementary 2D foundation models. Concretely, Chorus uses a shared 3D encoder over Gaussian primitives and lightweight per-teacher projectors to distill *(i) language-aligned semantics* from the SigLIP2 encoder [49], *(ii) generalist visual features* from DINOv3 [47], and *(iii) object-aware cues* from the Perception Encoder variant PE-Spatial [5, 25], which combines self-alignment with SAM-logit alignment to improve spatial locality while preserving semantics. Our multi-teacher design teaches the scene encoder breadth and complementarity, capturing high-level semantics, instance grouping, and fine spatial structure within a single 3D embedding space.

Chorus builds upon the "lift-then-align" paradigm established by SceneSplat [28], which lifts dense 2D language features to 3D Gaussians and uses them as pseudo-labels to train a feed-forward 3DGS encoder for open-vocabulary segmentation. However, SceneSplat's encoder was predominantly aligned with semantic information and demonstrated on semantic segmentation, leaving broader downstream applications (*e.g.*, instance grouping) and reasoning capabilities largely unexplored. Chorus generalizes this paradigm with multi-teacher pretraining to explicitly supervise with diverse signals in order to learn a versatile 3D feature representation. Our framework therefore results in 3D encoding that reaches superior performance across a diverse set of tasks, thereby producing *a holistic 3D scene encoder*.

We demonstrate the effectiveness of Chorus on the following tasks: semantic segmentation, open-vocabulary semantics, instance segmentation, open-vocabulary instances, and visual question answering. The evaluations are conducted on a comprehensive collection of datasets: ScanNet200 [44], ScanNet++ [61], Matterport3D [8], and our newly proposed 3DGS-native benchmark (with per-Gaussian labels) built upon InteriorGS [48]. In contrast, prior methods for generalizable 3D scene understanding typically specialize in a limited subset of tasks such as open-vocabulary tasks [27–29], semantics and instances [51, 53, 55], and VQA reasoning [9, 12, 14]. Our evaluation demonstrates that pretrained Chorus encoder can simultaneously serve as the most effective solution across this broad spectrum of scene understanding tasks. Furthermore, we carry out probing experiments (linear/decoder probing and full finetuning) for semantic and instance segmentation to assess the feature quality across the same datasets. In addition, we conduct data-efficiency studies that restrict supervision to limited scenes and sparse annotations, thereby stress-testing how much the pretrained encoder alone carries.

Besides 3DGS, we tested Chorus on several benchmarks that only support point clouds. For this purpose, we pretrained a new point-cloud-compatible 3D encoder using the Gaussians' centers, colors, estimated normals as the only inputs, while keeping all other training signals and losses identical. To our surprise, this variant is competitive with the recent self-supervised point cloud pretraining method Sonata [55] while using $\sim 39.9\times$ fewer training scenes. Chorus also exhibits favorable scaling as we move from subset to joint-dataset pretraining. These observations indicate that our multi-teacher pretraining successfully mines semantics, spatial locality, and instance grouping that carry over when the encoder is evaluated on point clouds tasks, despite the distribution difference. In practice, multi-teacher distillation over 3DGS is a practical, efficient route towards a general feed-forward scene encoder.

To further demonstrate versatility, Chorus facilitates out-of-domain adaptation by introducing a render-and-distill strategy that eliminates the need for costly 3D pseudo-label preprocessing. This approach leverages the inherent rendering capability of 3DGS: given a new dataset, we simply render 2D views, perform online teacher inference, and finetune our encoder with teacher knowledge. This makes the adaptation pipeline more lightweight and accessible.

Our contributions can be summarized as follows:

- A multi-teacher pretraining framework that aligns a native 3DGS encoder with diverse 2D teachers (language-aligned, generalist, and object-aware) via a shared backbone and per-teacher projectors.
- A holistic 3D scene encoding that produces highly structured and transferable embeddings for both 3DGS and PC inputs, leading to state-of-the-art performance on a broad range of tasks, while demonstrating data efficiency.
- A lightweight render-and-distill adaptation recipe that enables convenient out-of-domain finetuning without requiring costly 3D pseudo-label preprocessing.

2. Related Work

Self-supervised and Cross-modal Distillation for 3D. Self-supervised learning (SSL) has driven strong representation learning for 2D images [7, 15, 16] and has been actively explored for 3D data via contrastive and masked modeling [2, 34, 43, 56, 62]. Recently, Sonata [55] mitigates the geometric shortcut during point clouds self-supervised learning and [64] shows that joint 2D–3D SSL can yield more coherent spatial features than using a single modality. In parallel, knowledge distillation [18] has emerged as a powerful paradigm. Cross-modal distillation injects priors from 2D foundation models [26, 33, 40, 49, 63] into 3D, mitigating label scarcity and enabling semantic awareness in 3D representations [23, 36, 57, 58]. This distillation paradigm has progressed from single-teacher to multi-teacher aggregation in 2D domain, as shown by [17, 42, 45], which strengthens learning with complementary signals. We adopt this perspective and specialize it to 3DGS for the first time: instead of a single teacher or objective, Chorus distills from various 2D teachers (language-aligned, generalist, object-aware) to align embeddings with rich priors,

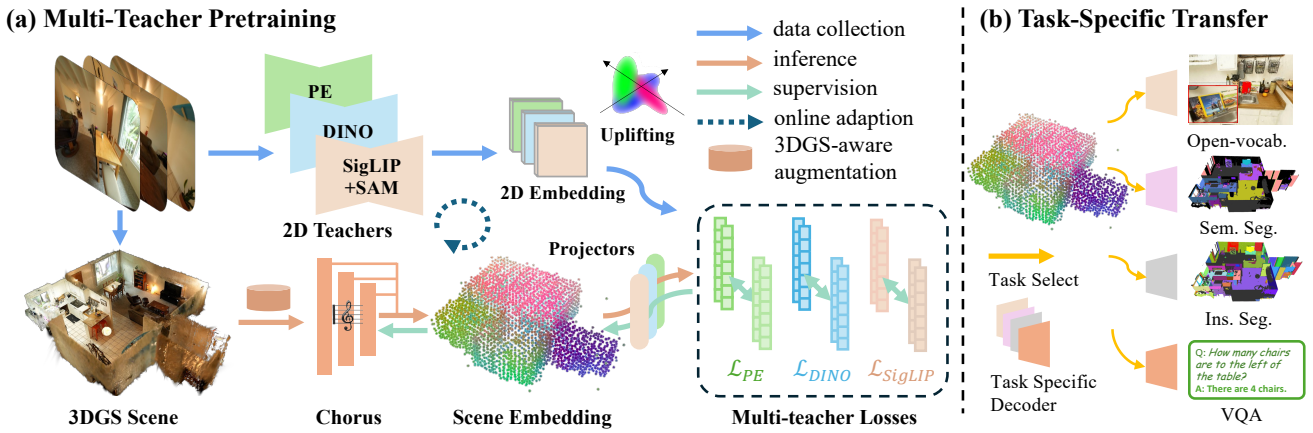

Figure 2. **Chorus Overview. (a) Multi-Teacher Pretraining.** We train a feed-forward 3DGS scene encoder to distill complementary signals–language-aligned (SigLIP), generalist (DINO), and object-aware (PE)–from 2D teachers. This knowledge is transferred into a shared embedding space via lightweight per-teacher projectors and losses. To accelerate out-of-domain adaptation, we support finetuning the encoder with online rendering-based supervision. **(b) Task-Specific Transfer.** Pretrained Chorus encoder enables diverse downstream tasks, including semantic and instance segmentation, open-vocabulary query, and 3D visual question answering (VQA).

while leveraging the inherent rendering capability of 3DGS.

3D Gaussian Splats Encoders. Unlike 3D point clouds, where representation learning is well explored [37, 38, 51, 53, 65], encoding 3D Gaussian Splats remains underexplored despite their richer parameter space that couples both appearance and geometry. ShapeSplat [30] pioneers object-level masked reconstruction for 3DGS objects, while Can3Tok [13] learns a scene-level VAE that tokenizes 3DGS scenes into latent codes. At the scene level, SceneSplat [28] lifts 2D semantic priors to train a generalizable 3DGS encoder for open-vocabulary semantics and introduces the SceneSplat-7K dataset. SceneSplat++ [29] further scales scene-level 3DGS data and establishes a comprehensive benchmark for language-aligned 3DGS methods. Building on this trajectory, Chorus proposes consolidating complementary 2D priors into a single feed-forward 3DGS encoder, producing holistic scene embeddings that transfer robustly across diverse tasks (semantic, instance, and question answering [14]).

3. Method

Chorus pretrains a general-purpose feed-forward Gaussian scene encoder by distilling knowledge from multiple 2D teachers. We first explain the pretraining data where the 2D feature maps are lifted to the 3DGS (§3.1). Then we present the multi-teacher framework, *i.e.*, a shared 3DGS encoder with lightweight per-teacher projectors and losses, including optional contrastive terms that exploit available semantic class/instance structure (§3.2). Next, we describe a rendering-based adaptation recipe that shortcuts adaptation via image-plane supervision, accelerating the out-of-distribution generalization (§3.3). Finally, we introduce our 3DGS-aware augmentations to aid pretraining (§3.4).

3.1. Lifting 2D Teachers for Supervision

3DGS scene rendering. A 3DGS scene is an optimized parameter set of N Gaussians:

$$\mathcal{G} = \{(\mathbf{x}_i, \mathbf{s}_i, \mathbf{q}_i, \alpha_i, \mathbf{c}_i)\}_{i=1}^{N} \tag{1}$$

to reproduce multi-view images via alpha composition and anisotropic Gaussians [22]. Each tuple contains parameters for a center $\mathbf{x}_i \in \mathbb{R}^3$, scale $\mathbf{s}_i \in \mathbb{R}^3_+$, orientation $\mathbf{q}_i \in \mathbb{H}$ (unit quaternion), opacity $\alpha_i \in [0,1]$, and color $\mathbf{c}_i \in [0,1]^3$. For a viewpoint p and a pixel $\mathbf{u} \in \mathbb{N}^2$, 3DGS renders colors as

$$\mathbf{C}(\mathbf{u}|p) = \sum_{i \in \mathcal{S}_{d,\mathbf{u}}} \underbrace{T_i\, \alpha_i(\mathbf{u}|p)}_{w_i(p,\mathbf{u})}\, \mathbf{c}_i, \quad T_i = \prod_{j<i} \big(1 - \alpha_j(\mathbf{u}|p)\big), \tag{2}$$

where $\mathcal{S}_{d,\mathbf{u}}$ is the depth-sorted set of splats intersecting the viewing ray.

Normalized uplifting. Let $F_{d,\mathbf{u}}$ be a 2D teacher feature at view p, pixel $\mathbf{u}$, and let f_i be the target feature on Gaussian i. Using the same rendering weights as in (2), we obtain uplifted supervision as a weighted average [31]:

$$f_i = \sum_{(p,\mathbf{u}) \in \mathcal{S}_i} \bar{w}_i(p,\mathbf{u})\, F_{p,\mathbf{u}}, \quad \bar{w}_i(p,\mathbf{u}) = \frac{w_i(p,\mathbf{u})}{\sum_{(p',\mathbf{u}') \in \mathcal{S}_i} w_i(p',\mathbf{u}')}, \tag{3}$$

where $\mathcal{S}_i$ are all view-pixel pairs contributing to feature f_i.

Teacher standardization. We supervise with three 2D teachers: SigLIP2 (*language-aligned*), DINOv3 (*generalist features*), and PE-Spatial (*object-aware*). Because teacher activations differ in scale/variance, we apply PHI-S [41], a PCA rotation followed by isotropic Hadamard scaling to achieve unit average per-channel variance while preserving cross-channel relationships. We denote $\widetilde{F}_{p,\mathbf{u}} = \text{PHI-S}(F_{p,\mathbf{u}})$ and use $\widetilde{f}_i$ analogously when needed.

3.2. Multi-Teacher Pretraining

Architecture. A shared 3DGS encoder g_θ maps Gaussian parameters to latent per-Gaussian features:

$$Z = g_\theta(\mathcal{G}) \in \mathbb{R}^{N\times d_z}. \tag{4}$$

Each teacher $\in \mathcal{T} = \{\text{lang}, \text{dino}, \text{pe}\}$ has a lightweight projector head h_t (2-layer MLP with LayerNorm and GELU) producing predictions $\hat{F}^{(t)} = h_t(Z) \in \mathbb{R}^{N\times d_t}$.

Per-teacher losses. For teacher t, we denote the uplifted supervision $\widetilde{F}^{(t)}$ and an optional validity mask $M^{(t)}$ (derived from feature norms / visibility). Our base matching loss combines cosine and smooth-ℓ_1 loss:

$$\mathcal{L}^{(t)}_{\text{match}} = \frac{1}{|M^{(t)}|}\sum_{i\in M^{(t)}} \lambda_1\Big(1-\cos\big(\hat{F}^{(t)}_i, \widetilde{F}^{(t)}_i\big)\Big) + \lambda_2\,\text{SmoothL1}\big(\hat{F}^{(t)}_i, \widetilde{F}^{(t)}_i\big), \tag{5}$$

for preserving both magnitude and angular alignment. We ℓ_2-normalize the inputs before calculating cosine terms.

Teacher-specific contrastive loss (optional). We add compact contrastive regularizers [28] when the source dataset provides semantic/instance labels.

- SigLIP2 teacher (semantic): pool class-wise means $\bar{F}^{(t)}_c = \text{mean}\{\hat{F}^{(t)}_i : i \in \mathcal{G}_c\}$, split each class into two disjoint halves A/B, and apply a bidirectional InfoNCE loss over ℓ_2-normalized $\bar{F}^{(t)}_{c,\{A,B\}}$ across classes.
- PE-Spatial teacher (instance): pool instance-wise means $\bar{F}^{(t)}_k = \text{mean}\{\hat{F}^{(t)}_i : i \in \mathcal{I}_k\}$ and similarly apply InfoNCE.

We write the loss term succinctly as $\mathcal{L}^{(t)}_{\text{con}}$ and put equations in the supplement.

Staged pretraining & total optimization objective. Teachers can start at different epochs. Let $\mathcal{A}(e) \subseteq \mathcal{T}$ denote the active set at training epoch e (*e.g.*, {lang, dino} from the start, then add pe). The total objective is

$$\mathcal{L}_{\text{total}}(e) = \sum_{t\in\mathcal{A}(e)} \lambda_t\Big(\mathcal{L}^{(t)}_{\text{match}} + \eta_t\,\mathcal{L}^{(t)}_{\text{con}}\Big), \tag{6}$$

with simple per-teacher weight λ_t and optional η_t. We empirically found that PHI-S standardization simplifies loss balancing, *i.e.*, equal weights of λ_t suffice across teachers.

3.3. Rendering-Based Adaptation

Given a novel data domain, we can adapt our pretrained encoder without precomputing 3D pseudo-labels by online inference. We sample camera poses $\{p\}$, conduct visibility culling on the input Gaussians, then run each 2D teacher on the rendered RGB to obtain feature maps $F^{(t)}_{p,\mathbf{u}}$, and obtain per-Gaussian predictions $\hat{F}^{(t)}_i$ with the current encoder and projector heads. Using the same compositing weights $w_i(p,\mathbf{u})$ as in Eq. (2), we render an inference feature map for each teacher t:

$$\hat{F}^{(t)}_{p,\mathbf{u}} = \sum_{i\in\mathcal{S}_{p,\mathbf{u}}} w_i(p,\mathbf{u})\,\hat{F}^{(t)}_i. \tag{7}$$

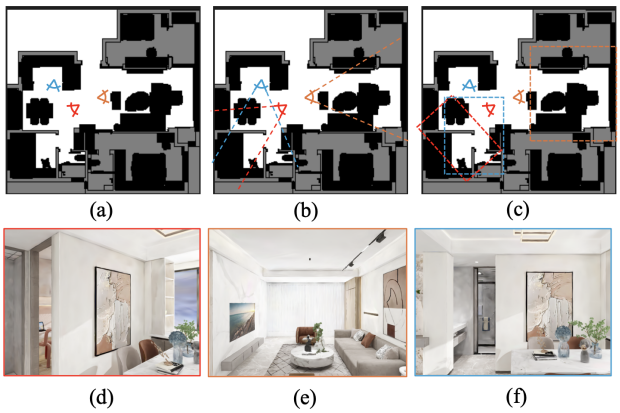

Figure 3. **Rendering-Based View Sampling and Pairing:** (a) Camera Location Sampling: We use Furthest Point Sampling to select camera positions that achieve broad spatial coverage across the entire *navigable* scene space. (b) Visibility Culling: For each location, we sample view angles and track the *visibility* of the 3D Gaussians across frames. (c) View Pairing and Selection: We obtain a minimum *2D bounding box* covering all visible Gaussians for a given view. Then candidate pairs of poses are calculated and sorted based on the overlap score. (d,e,f) Rendered images corresponding to the colored camera viewpoints.

Adaptation objective. Let Ω be the set of valid pixels with sufficient transmittance. We reuse the same per-teacher matching loss as in Eq. (5) (cosine + SmoothL1, with the same λ_1, λ_2), now applied to the 2D feature maps over Ω:

$$\mathcal{L}^{(t)}_{\text{img}} = \tfrac{1}{|\Omega|}\textstyle\sum_{(p,\mathbf{u})\in\Omega} \ell_{\text{match}}\big(\hat{F}^{(t)}_{p,\mathbf{u}},\ \widetilde{F}^{(t)}_{p,\mathbf{u}}\big). \tag{8}$$

This *render-and-distill* loop adapts Chorus using only rendered frames, accelerating the adaptation to new data.

View sampling and pairing pipeline. To enable adaptation on data without provided poses, we select informative views for 2D supervision in two stages: (1) *Informative view selection* – sample camera positions that are well distributed in the navigable space yet are neither too close to geometry nor heavily occluded; (2) *Contextual view pairing* – ensure that selected views share sufficient overlap to promote cross-view feature coherence. As illustrated in Fig. 3, we first sample camera positions proportional to the scene size to ensure coverage, and for each position generate eight candidate horizontal viewing directions. Directions whose center ray is too close to scene contents are discarded. For each valid view, we rasterize the Gaussians and record their visibility, then compute the minimum 2D axis-aligned bounding box enclosing all visible splats; only Gaussians that fall inside this region are kept as input for that training view. Finally, for each camera pose we sort the remaining poses by visibility overlap and form training pairs from high-overlap poses, ensuring sufficient multi-view context. Further details are provided in the supplement.

3.4. 3DGS-Aware Augmentations

Why point-cloud augmentations are suboptimal for 3DGS? Point clouds augmentations (dropout, elastic distortions, color/geometry jitter, etc.) are designed for i.i.d. sets of points whose attributes (position, color) are direct observations of 3D geometry/appearance. In contrast, a 3D Gaussian Splatting (3DGS) scene is an *optimized* parameter space. Naïve point-cloud jitter alters α_i and T_i in ways that are not motivated for splat-based rendering, and empirically we observe consistent performance drops when such jitter is applied to our encoder pretraining.

Design principle. We propose two augmentations that are *3DGS-aware*: (i) a *Rendering-Equivalent* perturbation, targeting for the augmented parameters that render approximately the same images, injects small, covariance-aware position noise primarily into low-opacity splats, and (ii) an *Immature-Manifold* perturbation to mimic earlier (blurrier) stages of optimization, selectively inflates per-splat covariances. Both augmentations are grounded in the rendering equation Eq. (2) and the observed 3DGS optimization dynamics [21, 24]; equations are provided in the supplement.

4. Experiments

We evaluate the pretrained Chorus encoder on diverse tasks. First, using the language-aligned projector, we report open-vocabulary semantic and instance segmentation. Next, we show that Chorus serves as an effective scene tokenizer for the language model, enabling question answering (§4.1). Later, we pretrain a point-cloud variant of Chorus using only Gaussians' center, color, estimated normal as inputs, which surprisingly competes with SoTA point clouds encoder consistently (§4.2). We analyze this unexpected robustness to gain understanding during ablation study (§4.3).

Implementation details. Our pretraining backbone adapts the 5-stage transformer encoder from Sonata [55] with the bottleneck feature dimension of 512. We employ the teachers of `SigLIP2-so400m-p16-512`, `DINOv3-ViTL16`, and `PE-Spatial-L14-448`. For pretraining data, we leverage the collected 3DGS scenes (center, color, opacity, quaternion, scale) from SceneSplat-7K [28], with our newly processed pseudo-labels for each teacher. We set teacher loss weights $\lambda_t = 1.0$ and balance the contrastive terms with $\eta_t = 0.02$ for $\mathcal{L}_{\text{con}}^{(t)}$. We pretrain the standard model (denoted ❄) using all 3DGS parameters as input, and a point-cloud variant (denoted ●) which instead uses center, color, estimated normal of the Gaussians while keeping all other settings the same. This variant is used for subsequent probing and finetuning to compare against point cloud encoders. For the rendering-based adaptation, we initialize with the pretrained Chorus encoder and for each batch we select 4 overlapping views. By default, the rendered image resolution is 480×640, and the rendered feature resolution is 120×160. We use bilinear interpolation to upsample the online encoded 2D teacher feature to match the rendered feature map. A learning rate of 2×10^{-4} is employed for the adaptation, which runs for 100 epochs. The SceneSplat-7K MP3D [8] subset was updated for Chorus training. Consequently, we re-trained [28] using public code for joint-data comparison. We refer to the supplement for additional training and experiments details.

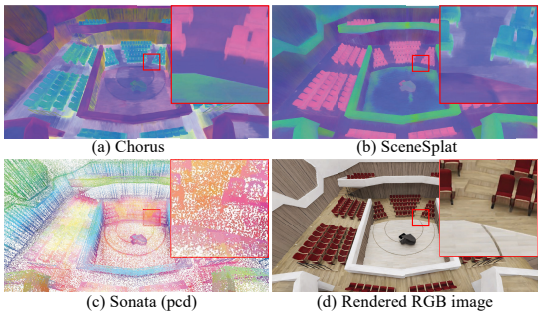

Figure 4. **Inference Feature PCA Visualization.** Features from different encoders on a concert hall. Chorus shows the best semantic consistency (see zoomed-in chairs and stairs in the back).

4.1. Main Results

Open-vocabulary semantic segmentation. Tab. 1 reports zero-shot semantic segmentation results on fine-grained ScanNet200, Matterport3D, ScanNet++, and InteriorGS benchmarks. Chorus achieves the best zero-shot performance, *e.g.*, compared to the previous SoTA SceneSplat, a 2.1% f-mIoU and 6.0% f-mAcc increase on the ScanNet200 benchmark after joint training, and when evaluated generalization on new data, a 5.7% f-mIoU and 5.8% f-mAcc increase on InteriorGS, while using $8.32\times$ less training data compared to the point clouds-based pretraining method [27], highlighting the efficiency of our 3DGS-based pretraining framework.

Open-vocabulary instance segmentation. We report open-vocabulary 3D instance segmentation on ScanNet200 in Tab. 2. The results confirm that our encoder's strong open-vocabulary semantic understanding translates to the instance level. Following the protocol from Mosaic3D [27], we adopt the instance proposals from Mask3D [46] for all baselines. Chorus achieves state-of-the-art performance among the methods that use 3D inputs only, outperforming prior point clouds-based open-vocabulary SoTA [27]. Notably, Chorus reaches a +7.6 mAP gain in recognizing the 66 tail classes, showing ability to recognize rare instances.

Rendering-based adaptation. As shown in Tab. 7, our lightweight adaptation recipe avoids heavy feature I/O during training and adds at most 0.1 s per view for on-the-fly feature rasterization—eliminating the ∼1 TB storage required to precompute teacher features for 800 training scenes. Its effectiveness is evident in Fig. 6: training on an additional 100 scenes from the InteriorGS dataset yields

Method	Training Source	#Training Scenes	ScanNet200 (200) f-mIoU	ScanNet200 (200) f-mAcc	Matterport3D (160) f-mIoU	Matterport3D (160) f-mAcc	ScanNet++ (100) f-mIoU	ScanNet++ (100) f-mAcc	❄ InteriorGS (72) f-mIoU	❄ InteriorGS (72) f-mAcc
OpenScene† [36]	SN	×1	6.4	12.2	5.7	10.7	8.8	14.7	–	–
PLA [11]	SN	–	1.8	3.1	–	–	–	–	–	–
RegionPLC [60]	SN	–	9.2	16.4	6.2	13.3	11.3	20.1	–	–
OV3D [20]	SN	–	8.7	–	–	–	–	–	–	–
Mosaic3D [27]	SN	–	13.0	24.5	8.6	17.8	16.2	27.1	3.8	8.2
❄ SceneSplat [28]	SN	–	18.9	31.7	10.8	18.7	14.7	24.7	6.1	8.5
❄ Chorus (ours)	SN	–	22.4	45.8	11.4	16.4	16.8	29.1	9.0	14.6
Mosaic3D [27]	SN, SN++, MP3D, ARKitS, S3D	×24.3	15.7	28.3	13.1	27.7	18.0	29.0	9.4	16.0
❄ SceneSplat [28]	SN, SN++, MP3D	×2.92	22.5	41.7	14.0	32.4	28.6	50.9	10.0	18.3
❄ Chorus (ours)	SN, SN++, MP3D	×2.92	**24.6**	**47.7**	**18.7**	**38.5**	**29.6**	**53.5**	**15.7**	**24.1**

Table 1. **Zero-Shot 3D Semantic Segmentation on the Fine-Grained ScanNet++ (100 classes) [61], Matterport3D (160 classes) [8], ScanNet200 (200 classes) [10] and InteriorGS (72 classes) [48] Benchmarks.** ❄ denotes 3DGS modality input. Chorus and SceneSplat [28] are the *only* methods that target 3DGS modality pretraining. We report the foreground mean IoU (f-mIoU) and foreground mean accuracy (f-mAcc) excluding wall, floor, ceiling classes, following [36, 60]. † denotes the official checkpoint and the baseline results are partly taken from [27]. Dataset abbreviations SN, SN++, ARKitS, MP3D, and S3D are short for ScanNet [10], ScanNet++ [61], ARKitScenes [4], Matterport3D [8] and Structured3D [66]. Chorus achieves noticeably better zero-shot performance, *e.g.*, a 3.2% f-mIoU and 9.0% f-mAcc increase on the ScanNet200, and when evaluated on new data, a 5.6% f-mIoU and 5.1% f-mAcc increase on InteriorGS compared to the previous SoTA SceneSplat, while using 8.32× less training data compared to the point clouds-based pretraining method [27].

Method	Inputs	3D Region Proposal Network	mAP 25	mAP 50	mAP head	mAP tail
Open3DIS	3D+2D	Superpoints + ISBNet + Grounded-SAM	32.8	29.4	27.8	21.8
SAI3D	3D+2D	Superpoints + SAM	24.1	18.8	12.1	16.2
OpenScene-3D	3D	Mask3D	7.2	6.2	10.6	0.7
RegionPLC	3D	Mask3D	9.7	8.6	15.6	1.7
OpenIns3D	3D	Mask3D	14.4	10.3	16.0	4.2
Mosaic3D	3D	Mask3D	17.8	16.0	**21.8**	5.4
❄ Chorus (ours)	3D	Mask3D	**19.6**	**18.0**	18.5	**13.0**

Table 2. **Open-Vocabulary 3D Instance Segmentation on ScanNet200.** Methods are grouped by input types, methods using both 3D+2D inputs requires expensive multi-view images processing, whereas Chorus is feed-forward and shows strength, especially on the 66 tail classes.

Methods	ScanQA EM1	ScanQA M	ScanQA R	Nr3D Sim	Nr3D M	Nr3D R
ScanQA [3]	–	13.1	33.3	–	–	–
3D-VLP [59]	–	13.5	34.5	–	–	–
Scene-LLM [12]	–	15.8	–	–	–	–
LL3DA [9]	14.3	22.8	34.7	48.1	5.8	9.9
GaussianVLM [14] ⊢ ❄ SceneSplat* [28]	14.4	**22.9**	34.8	48.2	20.8	19.2
GaussianVLM ⊢ ❄ Chorus (ours)	**14.8**	22.5	**37.4**	**50.6**	**22.5**	**28.8**

Table 3. **3D Scene Question and Answering.** Comparison across ScanQA (EM@1/M/R) and Nr3D (Sim/M/R).

a +2.7% mIoU gain under linear probing over our standard pretraining, indicating better domain adaptation. We also ablate teacher feature resolution during adaptation (Fig. 5), where even low-resolution 30×40 DINOv3 features produce a clear improvement, with further gains at higher resolutions and more adaptation scenes.

Language model-based question answering. We evaluate Chorus as the 3D encoder within an LLM-based pipeline for visual question answering and grounding (see Tab. 3), where swapping in Chorus yields consistent improvements on both benchmarks. Concretely, we follow GaussianVLM [14] and simply replace its 3D backbone: instead of using multi-level features from [28], we feed only the final Chorus encoder stage into the VLM, keeping all other components and training settings unchanged. We train and evaluate both the original GaussianVLM and Chorus-augmented variant on ScanQA [3] (3D-VQA) and Nr3D [1], on the metrics of EM1 (Top-1 Exact Match), M (METEOR), R (ROUGE), and Sim (Sentence Similarity). Fig. 7 provides qualitative VQA examples. As an additional benefit, leveraging only the last encoder stage of Chorus is lighter and faster, achieving approximately a 0.68× training time compared to GaussianVLM with SceneSplat.

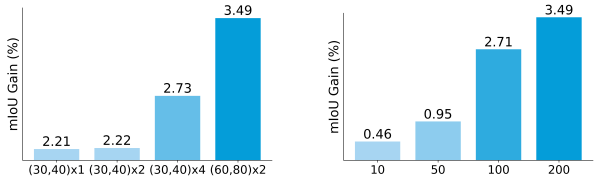

Figure 5. **2D Adaption Ablation.** Performance improves with higher teacher render resolution (left) and more adaptation scenes (right). The left x-axis denotes the 2D teacher's feature resolution, formatted as (feature size) × bilinear upsample factor.

4.2. Chorus on Point Clouds Tasks

Probing & finetuning of semantic segmentation. We evaluate the feature quality of our pretrained encoder via linear/decoder probing and full finetuning on five benchmarks, reported in Tab. 4. With only a learnable linear layer, Chorus can outperform the strong Sonata baseline across five benchmarks, *e.g.*, achieving mIoU gains on

Probing Exp.	ScanNet Val			ScanNet200 Val			ScanNet++ Val			Matterport3D (160)			✾ InteriorGS		
Methods	mIoU	mAcc	allAcc	mIoU	mAcc	allAcc	mIoU	mAcc	allAcc	mIoU	mAcc	allAcc	mIoU	mAcc	allAcc
MSC [52] (lin.)	21.8	32.2	65.5	3.3	5.5	57.5	8.1	11.9	64.7	–	–	–	–	–	–
Sonata [55] (lin.)	73.7	84.4	90.3	28.8	38.8	81.8	40.7	55.3	84.8	18.5	**25.8**	78.8	24.3	35.4	61.4
✾ / ● Chorus (lin.)	**75.2**	**84.8**	**90.5**	**36.0**	**47.2**	**82.8**	**48.8**	**63.2**	**86.4**	**20.0**	25.7	**79.4**	**27.0**	**37.2**	**62.6**
Sonata [55] (dec.)	**77.3**	**85.9**	**92.0**	30.1	39.4	**83.0**	46.6	58.9	**86.8**	19.0	26.2	79.4	27.2	38.6	65.3
✾ / ● Chorus (dec.)	75.0	83.1	90.6	**32.5**	**43.0**	82.2	**48.4**	**62.3**	86.7	**19.6**	**26.4**	**79.6**	**29.3**	**41.6**	**66.8**
PTv3 (sup) [53]	77.4	84.8	92.0	34.7	45.4	83.5	48.2	61.6	87.0	17.5	23.3	78.9	31.1	44.0	67.4
MSC (f.t.) [52]	78.2	85.3	92.2	33.4	43.7	83.4	48.7	61.9	87.2	–	–	–	–	–	–
Sonata (f.t.) [55]	78.6	86.6	92.3	34.4	44.0	84.0	49.9	60.7	**87.4**	21.3	27.6	80.2	30.7	41.8	66.2
✾ / ● Chorus (ours)	**79.4**	**87.7**	**92.4**	**40.9**	**52.3**	**84.1**	**52.9**	**66.2**	87.1	**23.6**	**31.0**	**80.5**	**31.8**	**43.3**	**68.9**

Table 4. **Semantic Segmentation Probing & Finetuning Experiments.** ✾ denotes 3DGS input and ● denotes point clouds input.

Data Efficiency	Limited Training Scenes				Limited Annotations			
Methods	1%	5%	10%	20%	20	50	100	200
Sonata [55] (lin.)	**45.3**	**62.3**	68.7	69.8	68.8	70.6	71.2	71.5
● Chorus (lin.)	42.0	60.3	**69.6**	**71.3**	**70.1**	**72.3**	**73.3**	**73.7**
Sonata [55] (dec)	**43.8**	**63.5**	69.5	**72.7**	69.4	**72.9**	**74.9**	**76.3**
● Chorus (dec.)	43.1	61.4	**69.7**	72.1	**70.8**	72.4	74.1	75.3
PTv3 [53] (sup.)	25.8	48.9	61.0	67.0	60.1	67.9	71.4	72.7
PPT [54] (sup.)	31.1	52.6	63.3	68.2	62.4	69.1	74.3	75.5
Sonata [55] (f.t.)	43.5	63.3	71.6	71.5	68.6	72.4	74.9	75.9
● Chorus (f.t.)	**43.9**	**64.0**	**73.9**	**75.0**	**73.1**	**76.1**	**77.2**	**77.4**

Table 5. **ScanNet Data-Efficient Benchmark.**

Methods	ScanNet Val		ScanNet200 Val		ScanNet++ Val	
	mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}
MSC [52] (lin.)	13.3	5.3	2.3	1.0	4.8	2.6
Sonata [55] (lin.)	**72.6**	**53.9**	30.0	20.9	33.5	24.5
● Chorus (lin.)	66.6	46.9	**31.6**	**21.9**	**37.0**	**27.9**
Sonata [55] (dec.)	**77.3**	**62.1**	36.2	29.3	39.4	33.5
● Chorus (dec.)	76.9	60.5	**38.8**	**31.8**	**41.9**	**33.8**
PTv3 [53] (sup.)	74.6	57.9	40.1	32.3	41.4	32.5
Sonata [55] (f.t.)	77.6	63.1	38.3	31.5	41.0	35.3
● Chorus (f.t.)	**78.4**	**63.4**	**39.3**	**33.7**	**42.9**	**37.2**

Table 6. **Instance Segmentation Probing and Finetuning.**

Supervise method	Preprocess (h)	Uplift (h)	Storage	Training Overhead
Uplifting	3.4	2.8	1080 GB	-
Rendering	0.2	0	8 GB	Rasterization (<0.1s/view)

Table 7. **Resource and Time Comparison of Uplifting-Based Supervision and Rendering-Based Adaptation.** Trade-offs between two approaches for training supervision on InteriorGS (800 scenes): uplifting based (preprocessing heavy) versus rendering-based adaptation (online computation heavy).

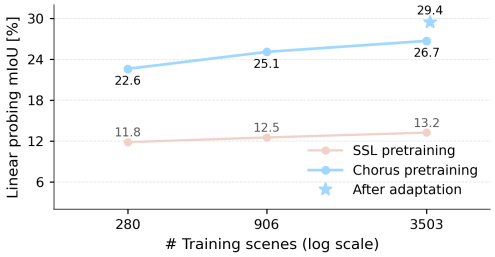

Figure 6. **Scaling Trend Together With Rendering-Based Adaptation.** Linear probing performance on InteriorGS *vs.* number of pretraining scenes. We compare our multi-teacher pretraining and the self-supervised pretraining [55] on 3DGS, Chorus scales faster and to higher accuracy. Our adaptation recipe yields a +2.7% mIoU gain on this new dataset using only 100 scenes.

ScanNet200 (36.0 *vs*. 28.8) and ScanNet++ (48.8 *vs*. 40.7). When fully finetuned, Chorus sets a new state-of-the-art on 4 out of 5 benchmarks, including ScanNet (79.4 mIoU) and ScanNet++ (50.2 mIoU). The advantage is particularly noticeable on ScanNet200, where Chorus achieves 40.9 mIoU, with a gain of +6.5 mIoU. Furthermore, Chorus consistently achieves relatively smaller gaps between linear probing and full finetuning. These results validate that our pretraining produces separable and semantic-aware features.

Probing & finetuning of instance segmentation. We extend analysis to instance segmentation in Tab. 6. Linear probing again shows the strength of our features; while Sonata leads on ScanNet, Chorus outperforms it on ScanNet200 (31.6 mAP_{25}) and ScanNet++ (37.0 mAP_{25}). When fully finetuned, Chorus remains competitive, achieving the best results on ScanNet++ (42.9 mAP_{25}) and performing comparably to top supervised methods on ScanNet.

Data efficiency experiments. We validate the benefit of our pretraining under data-scarce conditions on ScanNet in Tab. 5. The results show our encoder's pretrained features provide advantages over the Sonata baseline. When fully finetuned, Chorus consistently outperforms Sonata across all limited-scene (1%-20%) and limited-annotation (20-200 points/scene) settings. This demonstrates that our pretraining particularly helps in the low-data regime (*e.g.*, +4.5 mAP with 20 labels).

4.3. Ablation and Analysis

Why does Chorus work well on point clouds? We examine the Chorus variant that uses only Gaussians' centers, colors, normals as inputs while keeping the multi-teacher objectives unchanged. Despite the distribution gap between point clouds (observations) and 3DGS (optimized parameters), we posit two hypotheses: *(i)* 3DGS pretraining behaves like a *strong augmentation* of point clouds, induc-

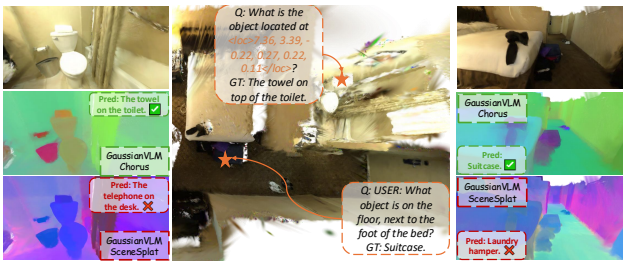

Figure 7. **VLM Qualitative Results.** We visualize a scene in ScanNet and object grounding (left) and QA results (right).

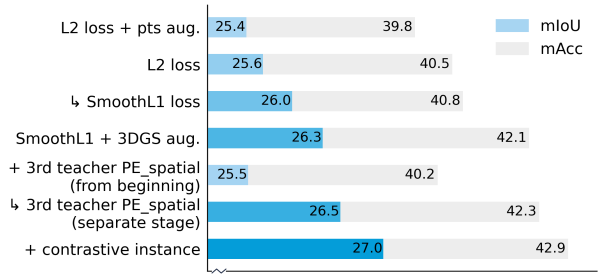

Figure 8. **Design Choice Ablation.** We validate the choices by evaluating zero-shot segmentation on ScanNet++ Val using a subset of training scenes. SmoothL1 loss, 3DGS-aware augmentations, introducing PE-Spatial in a separate stage, and an instance-level contrastive term each provide incremental gains.

Method	R@1 (PC→noisy PC)↑	Same-class@Incorrect top-1↑
Sonata	79.8%	75.0%
Chorus variant	**85.4%**	**78.0%**

Table 8. **Instance Retrieval From PC to Perturbed PC.** Averaged over 684 instances from 10 ScanNet++ Val scenes.

ing stable, noise-robust features; *(ii)* multi-teacher pretraining is more data-efficient than the self-supervised scheme, yielding better scaling.

To test *(i)*, we perform instance-level inference feature retrieval from original point clouds (PC) to perturbed PC (centers with Gaussian noise). We report R@1—fraction whose top-1 nearest feature is from the same instance—and *Same-class@Incorrect top-1*—when wrong, how often the prediction is at least the correct semantic class. The results are gathered from 684 instances using a subset of 10 scenes in the ScanNet++ Val split. Chorus variant is better on both (Tab. 8), indicating robustness to input noise. To test *(ii)*, we evaluate InteriorGS linear probing as the number of pretraining scenes grows (Fig. 6). When applying on 3DGS, Chorus pretraining scales faster than the self-supervised scheme used in Sonata. Taken together with Tab. 8, this suggests: (1) 3DGS-based pretraining induces noise-robust embeddings that transfer to point clouds, and (2) multi-teacher supervision supplies strong signals that keep improving with scale, contributing to the variant's strong PC performance despite the distribution gap. PCA analysis in Fig. 4 and supplement provide additional visualization.

Teachers ablation. We ablate the three teachers—SigLIP, DINO, and PE-Spatial—in two complementary views. In Tab. 9 we fix the language teacher and then add DINO, followed by PE; zero-shot semantic segmentation improves consistently on both the training dataset and InteriorGS as teachers are added, indicating complementary semantics (Lang) and general, object-aware structure (DINO/PE). We further ablate SigLIP teacher in the supplement. Together, there are non-redundant gains from each teacher.

Training source	Teachers			Val Split		InteriorGS	
	Lang	DINO	PE	$mIoU_{fg}$	$mAcc_{fg}$	$mIoU_{fg}$	$mAcc_{fg}$
ScanNet	✓	–	–	21.2	42.0	7.3	8.8
	✓	✓	–	**22.4**	**45.8**	**9.0**	**14.6**
ScanNet++ v2	✓	–	–	27.1	45.3	8.0	12.8
	✓	✓	–	29.4	55.8	9.3	16.0
	✓	✓	✓	**29.6**	**56.4**	**11.4**	**17.1**

Table 9. **Teachers Ablation with Zero-Shot Semantic Segmentation.** The "Teachers" columns mark included components (✓/–). We report foreground metrics likewise.

Design choice ablation. Fig. 8 evaluates training choices on ScanNet++ Val using a subset. SmoothL1 loss, 3DGS-aware augmentations, and an instance-level contrastive term each yield incremental improvements. A key finding is *when* to introduce PE-Spatial: staging it only in the second half of training outperforms enabling it from the start, suggesting that early PE may over-anchor to local features and compete with teacher alignment, whereas late applying refines spatial awareness after a stable backbone has formed.

Method	SceneSplat	Chorus	Sonata	Mosaic3D
#Model Params.	91.7M	131.3M	108.5M	39.1M
Inference Time/Scene	0.65s	0.70s	0.49s	0.25s

Table 10. **Model Size & Runtime.** Averaged on 100 scenes.

Runtime. Tab. 10 compares the model size and average inference time on 100 InteriorGS test scenes (with 965K Gaussians on average). Chorus has the slowest inference, but the runtime of 0.7s per scene is still practical.

5. Conclusion

We introduced Chorus, a multi-teacher pretraining framework that learns general-purpose 3D scene representations directly from 3D Gaussian splats. By aligning a native 3DGS encoder with complementary 2D foundation models, Chorus distills language-aligned, generalist, and spatially local cues into a unified 3D embedding that transfers well across scene understanding tasks. Extensive experiments on 3DGS-native and point clouds benchmarks show state-of-the-art performance and efficient render-and-distill adaptation to new domain. A remaining limitation is the offline cost of precomputing teacher pseudo-labels and an interesting direction is to move toward a unified point-cloud–3DGS encoder built on our findings.

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European conference on computer vision*, pages 422–440. Springer, 2020. 6

[2] Mohamed Afham, Isuru Dissanayake, Dinithi Dissanayake, Amaya Dharmasiri, Kanchana Thilakarathna, and Ranga Rodrigo. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9902–9912, 2022. 2

[3] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139, 2022. 6

[4] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021. 6

[5] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025. 2

[6] Ang Cao, Sergio Arnaud, Oleksandr Maksymets, Jianing Yang, Ayush Jain, Ada Martin, Vincent-Pierre Berges, Paul McVay, Ruslan Partsey, Aravind Rajeswaran, et al. From thousands to billions: 3d visual language grounding via render-supervised distillation from 2d vlms. In *Forty-second International Conference on Machine Learning*, 2025. 1

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2

[8] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 2, 5, 6

[9] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26428–26438, 2024. 2, 6

[10] 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*, pages 5828–5839, 2017. 6

[11] 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*, pages 7010–7019, 2023. 6

[12] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual reasoning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2195–2206. IEEE, 2025. 2, 6

[13] Quankai Gao, Iliyan Georgiev, Tuanfeng Y Wang, Krishna Kumar Singh, Ulrich Neumann, and Jae Shin Yoon. Can3tok: Canonical 3d tokenization and latent modeling of scene-level 3d gaussians. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9320–9331, 2025. 3

[14] Anna-Maria Halacheva, Jan-Nico Zaech, Xi Wang, Danda Pani Paudel, and Luc Van Gool. Gaussianvlm: Scene-centric 3d vision-language models using language-aligned gaussian splats for embodied reasoning and beyond. *arXiv preprint arXiv:2507.00886*, 2025. 2, 3, 6

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 2

[17] Greg Heinrich, Mike Ranzinger, Hongxu Yin, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. Radiov2. 5: Improved baselines for agglomerative vision foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22487–22497, 2025. 2

[18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[19] Haoyi Jiang, Liu Liu, Tianheng Cheng, Xinjie Wang, Tianwei Lin, Zhizhong Su, Wenyu Liu, and Xinggang Wang. Gausstr: Foundation model-aligned gaussian transformer for self-supervised 3d spatial understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11960–11970, 2025. 1

[20] Li Jiang, Shaoshuai Shi, and Bernt Schiele. Open-vocabulary 3d semantic segmentation with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21284–21294, 2024. 6

[21] Jaewoo Jung, Jisang Han, Honggyu An, Jiwon Kang, Seonghoon Park, and Seungryong Kim. Relaxing accurate initialization constraint for 3d gaussian splatting. 2024. 5

[22] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1, 3

[23] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19729–19739, 2023. 1, 2

[24] Shakiba Kheradmand, Daniel Rebain, Gopal Sharma, Weiwei Sun, Yang-Che Tseng, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. 3d gaussian splatting as markov chain monte carlo. *Advances in Neural Information Processing Systems*, 37:80965–80986, 2024. 5

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 2

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 2

[27] 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*, pages 14089–14101, 2025. 2, 5, 6

[28] Yue Li, Qi Ma, Runyi Yang, Huapeng Li, Mengjiao Ma, Bin Ren, Nikola Popovic, Nicu Sebe, Ender Konukoglu, Theo Gevers, et al. Scenesplat: Gaussian splatting-based scene understanding with vision-language pretraining. *arXiv preprint arXiv:2503.18052*, 2025. 1, 2, 3, 4, 5, 6

[29] Mengjiao Ma, Qi Ma, Yue Li, Jiahuan Cheng, Runyi Yang, Bin Ren, Nikola Popovic, Mingqiang Wei, Nicu Sebe, Luc Van Gool, et al. Scenesplat++: A large dataset and comprehensive benchmark for language gaussian splatting. In *NeurIPS*, 2025. 2, 3

[30] Qi Ma, Yue Li, Bin Ren, Nicu Sebe, Ender Konukoglu, Theo Gevers, Luc Van Gool, and Danda Pani Paudel. A large-scale dataset of gaussian splats and their self-supervised pretraining. In *2025 International Conference on 3D Vision (3DV)*, pages 145–155. IEEE, 2025. 3

[31] Juliette Marrie, Romain Ménégaux, Michael Arbel, Diane Larlus, and Julien Mairal. Ludvig: Learning-free uplifting of 2d visual features to gaussian splatting scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7440–7450, 2025. 1, 3

[32] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1

[33] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[34] 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. 2

[35] Qucheng Peng, Benjamin Planche, Zhongpai Gao, Meng Zheng, Anwesa Choudhuri, Terrence Chen, Chen Chen, and Ziyan Wu. 3d vision-language gaussian splatting. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[36] 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*, pages 815–824, 2023. 2, 6

[37] 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*, pages 652–660, 2017. 3

[38] 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, 2017. 3

[39] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20051–20060, 2024. 1

[40] 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*, pages 8748–8763. PmLR, 2021. 2

[41] Mike Ranzinger, Jon Barker, Greg Heinrich, Pavlo Molchanov, Bryan Catanzaro, and Andrew Tao. Phi-s: Distribution balancing for label-free multi-teacher distillation. *arXiv preprint arXiv:2410.01680*, 2024. 3

[42] Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12490–12500, 2024. 2

[43] Bin Ren, Guofeng Mei, Danda Pani Paudel, Weijie Wang, Yawei Li, Mengyuan Liu, Rita Cucchiara, Luc Van Gool, and Nicu Sebe. Bringing masked autoencoders explicit contrastive properties for point cloud self-supervised learning. In *ACCV*, 2024. 2

[44] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *European conference on computer vision*, pages 125–141. Springer, 2022. 2

[45] Mert Bülent Sarıyıldız, Philippe Weinzaepfel, Thomas Lucas, Pau de Jorge, Diane Larlus, and Yannis Kalantidis. Dune: Distilling a universal encoder from heterogeneous 2d and 3d teachers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30084–30094, 2025. 2

[46] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. *arXiv preprint arXiv:2210.03105*, 2022. 5

[47] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov,

Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 2

[48] Manycore Tech Inc. SpatialVerse Research Team. Interiorgs: A 3d gaussian splatting dataset of semantically labeled indoor scenes. https://huggingface.co/datasets/spatialverse/InteriorGS, 2025. 2, 6

[49] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 2

[50] Ziyi Wang, Yanran Zhang, Jie Zhou, and Jiwen Lu. Unipre3d: Unified pre-training of 3d point cloud models with cross-modal gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1319–1329, 2025. 1

[51] Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling, 2022. 2, 3

[52] Xiaoyang Wu, Xin Wen, Xihui Liu, and Hengshuang Zhao. Masked scene contrast: A scalable framework for unsupervised 3d representation learning. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 9415–9424, 2023. 7

[53] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4840–4851, 2024. 2, 3, 7

[54] Xiaoyang Wu, Zhuotao Tian, Xin Wen, Bohao Peng, Xihui Liu, Kaicheng Yu, and Hengshuang Zhao. Towards large-scale 3d representation learning with multi-dataset point prompt training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19551–19562, 2024. 7

[55] Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, and Julian Straub. Sonata: Self-supervised learning of reliable point representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22193–22204, 2025. 2, 5, 7

[56] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *European conference on computer vision*, pages 574–591. Springer, 2020. 2

[57] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*, pages 131–147. Springer, 2024. 2

[58] 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*, pages 1179–1189, 2023. 2

[59] Dejie Yang, Zhu Xu, Wentao Mo, Qingchao Chen, Siyuan Huang, and Yang Liu. 3d vision and language pre-training with large-scale synthetic data. *arXiv preprint arXiv:2407.06084*, 2024. 6

[60] 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*, pages 19823–19832, 2024. 6

[61] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 2, 6

[62] 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*, pages 19313–19322, 2022. 2

[63] 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*, pages 11975–11986, 2023. 2

[64] Yujia Zhang, Xiaoyang Wu, Yixing Lao, Chengyao Wang, Zhuotao Tian, Naiyan Wang, and Hengshuang Zhao. Concerto: Joint 2d-3d self-supervised learning emerges spatial representations. *arXiv preprint arXiv:2510.23607*, 2025. 2

[65] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16259–16268, 2021. 3

[66] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 519–535. Springer, 2020. 6

[67] Yuhang Zheng, Xiangyu Chen, Yupeng Zheng, Songen Gu, Runyi Yang, Bu Jin, Pengfei Li, Chengliang Zhong, Zengmao Wang, Lina Liu, et al. Gaussiangrasper: 3d language gaussian splatting for open-vocabulary robotic grasping. *arXiv preprint arXiv:2403.09637*, 2024. 1