

# Lexicon3D: Probing Visual Foundation Models for Complex 3D Scene Understanding

Yunze Man <sup>1</sup>Shuhong Zheng <sup>1</sup>Zhipeng Bao <sup>2</sup>Martial Hebert <sup>2</sup>Liang-Yan Gui <sup>1</sup>Yu-Xiong Wang <sup>1</sup><sup>1</sup> University of Illinois Urbana-Champaign    <sup>2</sup> Carnegie Mellon University<https://yunzeman.github.io/Lexicon3D>

## Abstract

Complex 3D scene understanding has gained increasing attention, with scene encoding strategies playing a crucial role in this success. However, the optimal scene encoding strategies for various scenarios remain unclear, particularly compared to their image-based counterparts. To address this issue, we present a comprehensive study that probes various visual encoding models for 3D scene understanding, identifying the strengths and limitations of each model across different scenarios. Our evaluation spans *seven* vision foundation encoders, including image-based, video-based, and 3D foundation models. We evaluate these models in *four* tasks: Vision-Language Scene Reasoning, Visual Grounding, Segmentation, and Registration, each focusing on different aspects of scene understanding. Our evaluations yield *key findings*: DINOv2 demonstrates superior performance, video models excel in object-level tasks, diffusion models benefit geometric tasks, and language-pretrained models show unexpected limitations in language-related tasks. These insights challenge some conventional understandings, provide novel perspectives on leveraging visual foundation models, and highlight the need for more flexible encoder selection in future vision-language and scene-understanding tasks.

## 1 Introduction

Recently, complex 3D scene understanding has emerged as a pivotal area in computer vision, encompassing tasks such as scene generation [24, 25, 26, 33, 74], reasoning [5, 35, 52, 55], and interaction [36, 108]. Leveraging large-scale vision foundation models, approaches like [42, 64, 68, 84, 91] have achieved promising results, thereby enabling a wide range of real-world applications, from autonomous driving [54, 75, 79, 112], robotics [57, 108], to multi-modal agents [1, 78].

While numerous studies [6, 67, 99] have provided guidance on the use of vision foundation models for 2D image-based tasks, the strategies for 3D scenarios remain unclear. A systematic understanding of complex real-world scenarios involves not only semantic and depth awareness [6], which is possible to evaluate within the 2D domain, but also geometric awareness and the ability to align with multi-modal information for reasoning and grounding tasks. To address this gap, our work evaluates the use of different types of visual foundation models for complex scene understanding and seeks to identify the strengths and limitations of each model in different scenarios. Ultimately, this study aims to contribute to the development of more effective and efficient scene understanding systems.

Concretely, we aim to address several key questions. First, given that most vision foundation models are trained on image or video data, we want to determine *whether 2D foundation models can effectively interpret 3D scenes*. Second, since video models inherently contain temporal information that captures aspects of the 3D structure as well, we investigate *whether they lead to better 3D feature*

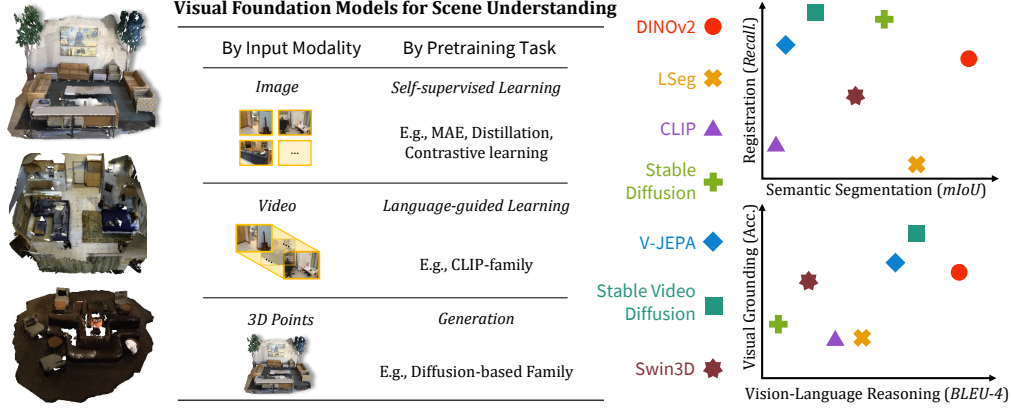


Figure 1: Evaluation settings and major results of different vision foundation models for complex 3D scene understanding. We probe visual foundation models of different input modalities and pretraining objectives, assessing their performance on multi-modal scene reasoning, grounding, segmentation, and registration tasks.

*representations compared to image models.* Finally, we seek to identify *the most suitable scenarios for different foundation models trained under various settings.*

To answer these questions, we design a unified paradigm to systematically probe visual encoding models for complex 3D scene understanding from different perspectives. Our evaluation spans *seven* vision foundation models in images, videos, and 3D-based ones, as shown in Table 1. Our evaluations are conducted among *four* diverse tasks: **Vision-Language Scene Reasoning** assesses the model’s ability to reason about scenes based on textual descriptions, evaluating *scene-level* representation; **Visual Grounding** tests the model’s capacity to associate language with specific objects within a scene, reflecting *object-level* representation; **Segmentation** evaluates the model’s ability to assign semantic labels to each pixel, assessing *semantic* understanding; **Registration** measures the performance of aligning different views of a scene, testing *geometric* capacity. Through these tasks, our aim is to explore the strengths and weaknesses of different vision foundation models in 3D scene understanding, providing insights into their applicability in various scenarios. With the major results demonstrated in Figure 1, our key findings include:

- Image or video foundation models achieve promising results for 3D scene understanding. Among them, DINOv2 [58] demonstrates the best overall performance, showing strong generalizability and flexibility, which is consistent with the observation in 2D [6]. Our evaluation further verifies its capability in global and object-level 3D vision-language tasks. It can serve as a general backbone for 3D scene understanding.
- Video models, benefiting from temporally continuous input frames, excel in object-level and geometric understanding tasks by distinguishing instances of the same semantics in a scene.
- Visual encoders pretrained with language guidance *do not* necessarily perform well in other language-related evaluation tasks, challenging the common practice of using such models as default encoders for vision-language reasoning tasks.
- Generative pretrained models, beyond their well-known semantic capacity, also excel in geometrical understanding, offering new possibilities for scene understanding.

We name our work as **Lexicon3D**, a unified probing architecture and the first comprehensive evaluation of 3D scene understanding with visual foundation models. The key findings we have achieved above, in conjunction with other interesting observations, suggest exploring more flexible encoder selections in future vision-language tasks to optimize performance and generalization.

## 2 Related Work

Our work is closely related to methods that focus on extraction of features from images, videos, and 3D assets, as well as learning joint spaces for vision-language fusion. A large body of recent literature has explored the representation learning for multi-modal visual inputs and their complementary performance in image understanding. In contrast, our paper presents a comprehensive analysis of the use of pretrained visual encoders for *zero-shot* 3D scene understanding. *To the best of our*

Model	Input Modality	Architecture	Supervision	Dataset
DINOv2 [58]	Image	ViT-L/14	SSL	LVD-142M
LSeg [44]		ViT-L/16	VLM	LSeg-7Mix
CLIP [65]		ViT-L/14	VLM	WIT-400M
Stable Diffusion [70]		UNet	Generation	LAION
V-JEPA [11]	Video	ViT-L/16	SSL	VideoMix2M
Stable Video Diffusion [12]		UNet	Generation	LVD-F
Swin3D [93]	3D Points	Swin3D-L	Annotation	Structure3D

Table 1: Details of the seven evaluated visual encoding models.

*knowledge, we are the first to examine pretrained video encoders on 3D scene understanding tasks and to compare image, video, and 3D point encoding strategies in this context.*

**Image self-supervised learning.** In recent years, learning robust and generalizable pretrained image representations has become a prevalent research direction in computer vision and multi-modal research. One line of work focuses on learning task-agnostic image features using self-supervised learning (SSL) signals, which include pretext tasks such as colorization [100], inpainting [62], transformation prediction [27], and self-distillation [14, 18, 19, 29, 30]. The recent development of the patch-based image tokenizer, ViT [22], has also led to the emergence of mask autoencoder architectures (MAE) for feature extraction [8, 31, 111]. Of particular interest, DINOv2 [58], combining a masked-image modeling loss and an invariance-based self-distillation loss, has become one of the most scalable and competitive self-supervised learning architectures using only image signals. Another line of work proposes learning image features with text guidance, *i.e.*, using textual descriptions to guide the pretraining of the image encoders [38, 53]. Building upon the powerful image-text encoder CLIP [65], LSeg [44] and BLIP [45, 46] extend the image pretraining objective to more complex visual perception tasks by incorporating pixel-level semantic understanding and encouraging better alignment with large language models (LLMs) [13, 66, 102, 103], respectively.

**Video and 3D representation learning.** Self-supervised representation learning has also been explored in the context of videos and 3D point clouds. Extending the success of the CLIP architecture [65] from images to videos, a body of work proposes to pretrain a video encoder by aligning the feature space with textual guidance extracted from video captions [3, 85, 89, 97]. Other pretext tasks used in video representation learning include next frame prediction [10] and MAE [28, 80, 83]. Among them, Bardes et al. [11] adapt the MAE-inspired joint embedding prediction architecture (JEPA) [4, 43] to the spatio-temporal domain, achieving state-of-the-art performance on a wide spectrum of video and image tasks. Despite the extensive research on 2D visual foundation encoders, pretrained models for 3D point clouds are significantly fewer due to the lack of large-scale 3D datasets. Existing work has explored contrastive pretraining [37, 88, 105] and masked signal modeling [48, 59, 87, 92, 96, 101] for point representation learning. Recently, benefiting from the rapid advancement of 3D data rendering and a large synthetic dataset [109], Swin3D [93] has outperformed other pretraining methods by a significant margin using supervised pretraining.

**Generation and mixture of experts (MoE) for feature extraction.** With the success of diffusion-based generative models [32, 70, 76], a line of research has begun to explore their role in image perception tasks. These methods extract feature maps or attention maps of a given image from the U-Net architectures of diffusion models and perform various downstream tasks, including depth estimation [23, 71, 107], semantic segmentation [9, 51, 56, 86, 107], object detection [17], and panoptic segmentation [90]. Another line of work [60, 98, 99] investigates the complementary nature of different embeddings extracted by multiple foundation backbones and their joint effect on downstream tasks [6, 67]. However, these investigations have been limited to the 2D domain, leaving the potential of leveraging pretrained encoders for perception and reasoning tasks in complex 3D scenes [5, 21, 34, 35, 40, 52, 55, 63, 113] largely unexplored.

### 3 Probing Visual Encoders for Scene Understanding

The objective of our Lexicon3D is to evaluate different visual foundation models in complex scene understanding tasks. We first construct a unified architecture capable of probing different visual foundation models on a spectrum of downstream tasks. Then, we break down the 3D scene understanding

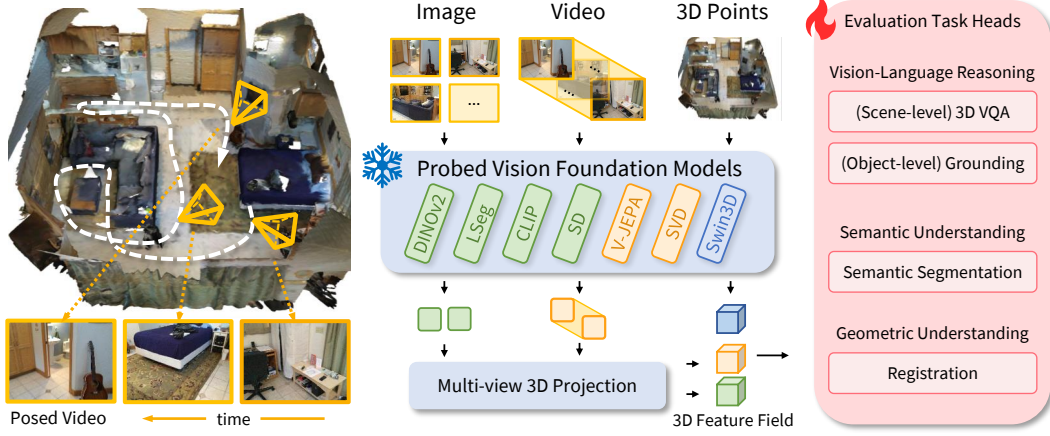


Figure 2: **Our unified probing framework** to evaluate visual encoding models on various tasks.

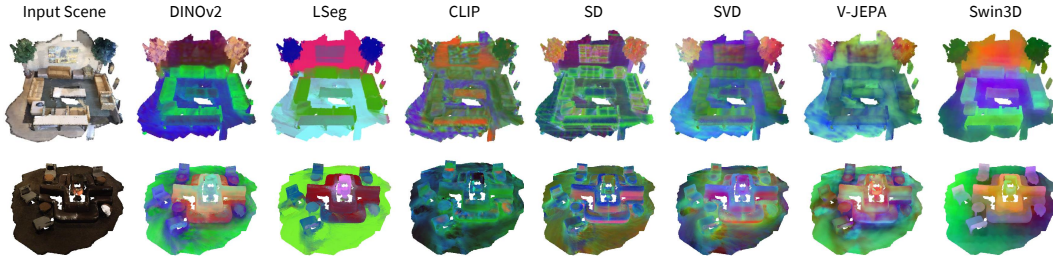


Figure 3: **Visualizations of extracted scene features** from different visual encoders using PCA. The clear distinction between colors and patterns demonstrates the behaviors of different models.

task into four sub-tasks, including (1) vision-language reasoning, (2) visual grounding, (3) semantic understanding, and (4) geometric understanding, for a more detailed evaluation.

### 3.1 A Unified Probing Framework

We design a unified framework, as shown in Figure 2, to extract features from different foundation models, construct a 3D feature embedding as scene embeddings, and evaluate them on multiple downstream tasks. For a complex indoor scene, existing work usually represents it with a combination of 2D and 3D modalities. For realistic scenarios [15, 20, 94], videos are usually first captured with handheld cameras and then 3D points are obtained from reconstruction algorithms such as COLMAP [72]. For digital and synthetic scenarios [69, 109], 3D assets are designed and generated first, before images and/or videos are rendered within the created space. Given a complex scene represented in posed images, videos, and 3D point clouds, we extract their feature embeddings with a collection of vision foundation models. For image and video-based models, we project their features into the 3D space for subsequent 3D scene evaluation tasks with a *multi-view 3D projection module*. Following [21, 34, 35, 63], for a point cloud  $\mathbf{P}$ , this module produces features  $f_{\mathbf{p}}$  for each point  $\mathbf{p} \in \mathbf{P}$  given image features  $f$  and the pose and camera information  $\mathbf{K}, \mathbf{R}$ . We first project all points onto the image plane to obtain their corresponding pixel features. Concretely, for a point  $\mathbf{p}$ , we obtain its projected pixel  $\mathbf{u}$  on the image  $i$  with

$$\tilde{\mathbf{u}} = \mathbf{K}_i \mathbf{R}_i \tilde{\mathbf{p}}, \quad \tilde{\mathbf{u}}, \tilde{\mathbf{p}} \text{ represent homogeneous coordinates of } \mathbf{u}, \mathbf{p}. \quad (1)$$

In addition, we use an indicator function  $\mathcal{I}(\mathbf{p}, i)$  to represent whether a point  $\mathbf{p}$  is visible in the image of the  $i$ -th frame. After finding corresponding pixels of the given point in all image frames, we use mean pooling as an aggregation function  $\phi$  to fuse all pixel features to form the point feature  $f_{\mathbf{p}}$ . Assuming there are  $M$  images in total, the projection and aggregation process is represented as:

$$f_{\mathbf{p}} = \phi_{i=1}^M (\mathcal{I}(\mathbf{p}, i) \cdot f_i(\mathbf{K}_i \mathbf{R}_i \tilde{\mathbf{p}})). \quad (2)$$

After projection, we obtain 3D feature fields represented as point cloud feature embeddings for each visual foundation model. We use these as input to the shallow probing heads to evaluate various

Model	ScanQA (higher means better for all metrics)					SQA3D (higher means better for all metrics)				
	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr	EM-1	BLEU-1	METEOR	ROUGE	CIDEr
3D-LLM [35] (for ref.)	39.3	12.0	14.5	35.7	69.4	48.1	47.3	35.2	48.6	124.5
DINOv2	39.2	13.4	15.3	36.8	73.2	50.1	49.5	35.6	50.7	129.1
LSeg	36.8	11.5	14.6	36.0	71.0	47.4	46.5	33.2	47.8	122.5
CLIP	36.4	10.7	14.4	36.0	70.3	48.1	47.3	34.6	48.6	124.5
StableDiffusion	35.5	11.7	14.1	34.9	68.2	47.7	47.2	33.6	48.3	124.0
V-JEPA	37.4	12.1	14.7	36.7	71.4	48.4	48.1	34.8	50.0	125.7
StableVideoDiffusion	38.5	12.5	14.5	35.4	70.6	48.5	47.9	34.4	49.0	127.7
Swin3D	36.1	10.5	13.9	35.4	70.0	48.3	48.0	34.1	47.3	123.9

Table 2: Evaluation of vision-language reasoning on ScanQA [5] and SQA3D [52] datasets. Top-2 results for each metric are shown in red and green, respectively. Prior method 3D-LLM results are shown for reference, indicating relative position of our evaluation results with respect to leading models trained on this task.

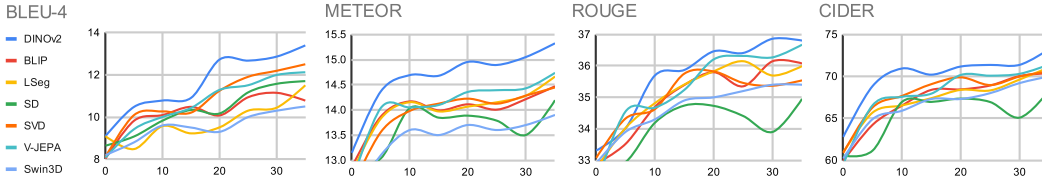


Figure 4: Evaluation curves on the ScanQA benchmark. DINOv2 exhibits clear superior performance.

downstream tasks. To minimize the effect of the model finetuning process, we freeze the parameters for the encoding models to be evaluated, and only tune the linear or shallow probing head for all tasks.

**Models.** In this work, we primarily focus on evaluating visual foundation models that are frequently leveraged by recent complex scene understanding and multi-modal reasoning models. A complex scene can often be represented in posed 2D images and videos or in 3D point clouds. The image and video modalities sacrifice explicit geometry information, but they preserve rich and dense semantic and textural information of a scene. Conversely, the point cloud modality offers the opposite trade-offs. Additionally, the 2D modalities benefit from strong foundation models trained on vast amounts of data, while 3D point backbones only leverage much smaller datasets.

We categorize visual foundation models into three categories, with an overview of the evaluated models provided in Table 1. For image encoders, we evaluated DINOv2 [58], LSeg [44], CLIP [65], and StableDiffusion (SD) [70]. For the video modality, we evaluate V-JEPA [11], the state-of-the-art video understanding model succeeding VideoMAE [80, 83] for a wide spectrum of perception and reasoning tasks, as well as StableVideoDiffusion (SVD) [12], a video generative model. The lack of large-scale 3D scene-level datasets hinders the development of strong zero-shot generalizable 3D foundation models as opposed to their 2D counterparts. However, for comparison, we evaluated Swin3D [93], a 3D backbone that achieves leading performance in zero-shot perception tasks in multiple evaluation datasets compared to earlier methods [37, 88, 105]. Swin3D is pretrained on Structured3D [109], a dataset 10 times larger than ScanNet [20].

**Feature visualization.** Figure 3 visualizes the scene features extracted by the vision foundation models. To visualize a high-dimensional feature space with  $C$  channels, we apply principal component analysis (PCA) to reduce the feature dimensions to three, normalize them to range  $[0, 1]$ , and interpret them as RGB color channels. The visualizations reveal several intuitive findings. The image models, DINOv2 and LSeg, demonstrate strong semantic understanding, with LSeg exhibiting clearer discrimination due to its pixel-level language semantic guidance. The diffusion-based models, SD and SVD, in addition to their semantic modeling, excel at preserving the local geometry and textures of the scenes, because of the generation-guided pretraining. The video models, SVD and V-JEPA, showcase a unique ability to identify different instances of the same semantic concepts, such as the two trees in the first scene and the chairs in both scenes. The 3D model, Swin3D, also exhibits strong semantic understanding. However, due to limited training data and domain shift, its quality is not on par with the image foundation models, despite being pretrained on perfect semantic annotations.

### 3.2 Vision-Language Reasoning

The vision-language reasoning task requires a model to engage in dialogues or answer questions about global understanding and local concepts and objects related to a given complex 3D indoor



scene. Following [35, 108], we formulate this as a visual-question answering (VQA) task using Large Language Models (LLMs) as the backbone – given a 3D scene from multi-view images and point clouds, and a user-prompt question, the LLMs are asked to generate the answer to the question in an auto-regressive way. This task encompasses universal language-guided reasoning of the complex indoor scene, ranging from global layout to local details.

**Datasets and optimization.** We evaluate the performance on two challenging indoor 3D VQA datasets: ScanQA [5] and SQA3D [52]. Following the evaluation methodology of [5, 35, 52, 55], we report the metrics BLEU [61], ROUGE [47], METEOR [7], and CIDEr [82]. We finetune a Q-Former module [46] to align features from different encoders to the LLM input space. More datasets and optimization details are provided in the supplementary material.

**Evaluation results.** Table 2 and Figure 4 present the results of our evaluation. We observe that image and video encoders generally outperform the 3D point encoder, with DINOv2 achieving the best performance, followed closely by V-JEPA and SVD. Interestingly, we find that for LSeg and CLIP, which are pretrained by language guidance, *their advantage in language alignment does not translate into superior performance on the LLM-guided VQA task*. This finding challenges the common practice of using language-pretrained visual foundation models [44, 45, 46, 65] as default encoders for LLM-based vision-language reasoning tasks. Instead, it suggests the importance of considering a wider range of encoders, such as DINOv2 and V-JEPA, to support such models.

### 3.3 Visual Grounding

Visual grounding is the task of locating an object in a 3D scene based on a text description. Compared to the 3D VQA task, visual grounding places a greater emphasis on object-level reasoning and matching capabilities. The task can be broken down into two sub-tasks: object detection and target discrimination (matching the text description with the target object). Although some methods focus on learning models to tackle both tasks [16, 104], others primarily focus on the discrimination problem [2] by assuming access to ground-truth bounding boxes. For simplicity and to prevent task entanglement, we adopt the latter setting in our evaluation. More specifically, given a 3D scene in the form of multi-view images and point clouds, a free-form language description of objects, and the ground-truth 3D bounding boxes of all objects in the scene, our model’s objective is to find the correct objects in the scene that match the language description. We believe that the object detection task requires semantic information from the visual encoder, which is similar in nature to the semantic segmentation task and will be analyzed in Section 3.4.

For the target discrimination task, we first obtain the feature for each object in the scene by taking the average pooling of all points inside its ground truth bounding box. Following Multi3DRefer [104], we use a CLIP text encoder to tokenize the text description, and adopt the attention head in [104] to fuse the text and visual embeddings from the previous steps and output an object score.

**Datasets.** We evaluate on the ScanRefer [16] dataset, which provides 51K text descriptions of 11K objects in 800 ScanNet scenes [20]. We report accuracy for *unique*, *multiple*, and *overall* categories, with *unique* referring to instances that have a unique semantic class in a given scene (easier).

**Optimization.** The model is trained with a cross-entropy loss using the AdamW [50] optimizer following [104]. We train our models for 30 epochs until convergence.

**Evaluation results.** Table 3 presents our results, which show that video encoding models demonstrate significant advantages over image and 3D encoders. The performance gap primarily lies in the *multiple* category, indicating that these models excel at discriminating the correct object among multiple objects of the same semantic category. This capability largely stems from the temporally continuous input frames, which provide instance-aware multi-view consistent guidance. In comparison, the image encoder LSeg, with its language-guided pretraining features aligned with language semantics, can also achieve high accuracy in the *unique* category. However, its performance drops significantly in the *multiple* category.

Model	Unique ↑	Multiple ↑	Overall ↑
M3DRef [104] ( <i>for ref.</i> )	88.0	46.1	54.3
DINOv2	87.0	43.4	52.0
LSeg	88.1	41.2	50.4
CLIP	86.5	41.6	50.4
StableDiffusion	86.4	41.9	50.6
V-JEPA	85.6	44.9	52.9
StableVideoDiffusion	88.0	46.5	54.7
Swin3D	85.7	43.2	51.6

Table 3: Evaluation of 3D object grounding on ScanRefer [16]. Video models exhibit significant advantages.

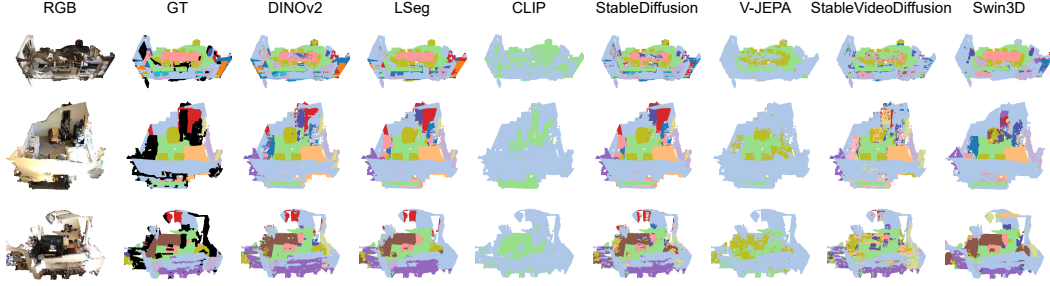


Figure 5: Visualization of 3D semantic segmentation on ScanNet [20]. Image encoders obtain better performance.

**Insights from vision-language tasks.** Our evaluation of vision-language reasoning and visual grounding reveals several key findings: (1) The DINOv2 unsupervised image learning model demonstrates strong generalizability and flexibility in global and object-level vision-language tasks. (2) Video encoders benefit from temporally continuous input frames and learn to distinguish instances of the same semantics in a scene, which is highly valuable for object-level understanding tasks. (3) Visual encoders pretrained with language guidance do not necessarily lead to strong performance in other language-related evaluation tasks. These findings suggest exploring a more flexible encoder selection in future vision-language tasks to optimize performance and generalization.

### 3.4 Semantic Segmentation

Semantic segmentation is the task of predicting semantic labels at each 3D position, which requires fine-grained semantic awareness of the scenes. As mentioned in Section 3.1, all types of features are unified in the form of point clouds; therefore, semantic labels are predicted for each point within the point cloud in our setting. More specifically, given a 3D scene in the form of multi-view images and point clouds, the objective in this task is to predict the semantic label for every point in the cloud.

**Dataset.** We conduct the experiments on the ScanNet [20] segmentation dataset which has 1,201 and 312 scenes for training and validation, respectively, with a total of 20 semantic classes for evaluation.

**Optimization.** To make the semantic prediction performance better reflect the fine-grained semantic understanding capability of different features, we use a single linear layer followed by a Sigmoid function to perform a linear probe to predict the probability distribution  $\mathbf{y} \in \mathbb{R}^{N \times C}$  for all the labels from the foundation model feature  $\mathbf{x} \in \mathbb{R}^{N \times d}$ :  $\mathbf{y} = \text{Sigmoid}(\text{FC}(\mathbf{x}))$ , where  $N$  is the number of points in each point cloud,  $d$  is the feature dimension, and  $C$  is the number of classes for segmentation. We adopt the standard Adam optimizer [41] with a learning rate of 1e-4 and use a cross-entropy loss to train the linear layer for 20 epochs.

**Evaluation results.** Table 4 and Figure 5 demonstrates that image encoders have better performance than video and 3D encoders on 3D semantic segmentation tasks. The reason is that image encoders like DINOv2 and LSeg gain their semantic awareness during training with contrastive objectives via either SSL or language-driven guidance. In comparison, video encoders have the risk of over-smoothing the multi-view information during multi-frame integration, which may harm the fine-grained semantic understanding capability. As for 3D encoders like Swin3D, the data scarcity in 3D compared to 2D for training the foundation models leads to inferior performance on semantic understanding.

Model	Acc $\uparrow$	mAcc $\uparrow$	mIoU $\uparrow$
GrowSP [106] (for ref.)	73.5	42.6	31.6
DINOv2	82.5	75.4	62.8
LSeg	78.2	58.5	47.5
CLIP	39.7	7.2	3.4
StableDiffusion	77.2	55.5	42.6
V-JEPA	58.7	13.2	8.1
StableVideoDiffusion	71.5	40.5	30.4
Swin3D	78.0	44.8	35.2

Table 4: Results of semantic segmentation.

### 3.5 Registration: Geometric Correspondence

To evaluate the geometric information contained in the foundation model features, we design the following new task, **partial scene registration**, based on the point cloud registration [49, 95] task. From a complete point cloud representing the entire scene, we sample a pair of point clouds  $P_1 \in \mathbb{R}^{N_1 \times 3}$  and  $P_2 \in \mathbb{R}^{N_2 \times 3}$  within the scene, where  $P_1$  and  $P_2$  contain all the points that can be observed in two sets of consecutive views, respectively. Our goal is to find the homography matrix  $H$

Model	RR@0.05m (%)	RR@0.1m (%)	RR@0.2m (%)	RRE (°)	RTE (m)
DINOv2	82.1	93.9	96.8	1.72	0.14
LSeg	4.8	23.7	63.8	9.80	0.59
CLIP	18.6	51.3	78.2	7.96	0.44
StableDiffusion	91.7	96.8	98.4	1.15	0.09
V-JEPA	90.4	96.5	99.4	1.37	0.10
StableVideoDiffusion	96.8	99.0	99.7	0.83	0.06
Swin3D	60.3	81.1	91.3	3.60	0.23

Table 5: Evaluation of partial scene registration on ScanNet [20]. We employ Registration Recall (RR) at various RMSE thresholds, Relative Rotation Error (RRE), and Relative Translation Error (RTE) as evaluation metrics. A higher RR indicates better performance, while lower RRE and RTE values signify superior results.

Model	Time (sample)	Time (scene)	Mem.
DINOv2	25.0 ms	7.5 sec	1.19 G
LSeg	291.2 ms	87.4 sec	2.51 G
CLIP	34.5 ms	10.4 sec	1.19 G
StableDiffusion	42.7 ms	12.8 sec	5.08 G
V-JEPA	175.1 ms	3.3 sec	1.31 G
StableVideoDiffusion	667.1 ms	12.5 sec	11.70 G
Swin3D	937.4 ms	0.9 sec	1.34 G

Table 6: Complexity analysis of visual foundation models.

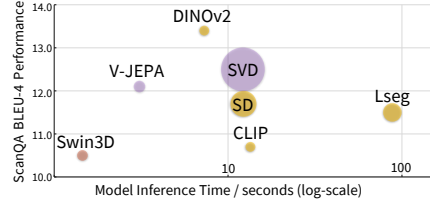


Figure 6: Memory usage of different encoders. An ideal model should be a small circle and be positioned in the upper left.

that correctly transforms the points in  $P_1$  to register with  $P_2$ . Compared to the semantic segmentation task evaluated in Section 3.4, the partial scene registration task requires the foundation model features to have the capability of finding *geometric correspondence* for registration, which cannot be achieved simply by finding the correspondence according to semantic understanding. For example, in semantic correspondence, we may find two semantically similar points, one on the left side of the sofa in  $P_1$ , while the other on the right side of the sofa in  $P_2$ . As a result, if we register the two partial point clouds solely based on semantic correspondence, we will fail to find the correct homography to align one point cloud with the other. The foundation model features need to be equipped with geometric understanding capability to achieve decent performance on our partial scene registration task.

**Dataset and probe head.** We build our partial scene registration benchmark based on the ScanNet [20] dataset. For each scene in ScanNet, we choose views #0 ~ #31 and views #32 ~ #63 for rendering  $P_1$  and  $P_2$ , respectively, so that they can have a certain level of overlap that allows the registration of two partial point clouds. Afterwards,  $P_2$  is transformed by a homography  $H$  that consists of a rotation  $\mathbf{R} \in \text{SO}(3)$  and a translation  $\mathbf{t} \in \mathbb{R}^3$ .  $\mathbf{R}$  is created by a randomly generated quaternion  $\mathbf{q} \in \mathbb{R}^4$  for each scene, while each component of  $\mathbf{t}$  is randomly sampled from the uniform distribution  $[-1.0\text{m}, 1.0\text{m}]$ . We follow REGTR [95] to adopt a transformer cross-encoder module, followed by a lightweight decoder, to obtain the corresponding position of each point in the other point cloud. More details on dataset and optimization are provided in the supplementary material.

**Evaluation results.** Table 5 demonstrates the results for the partial scene registration. We can observe that StableDiffusion and StableVideoDiffusion showcase superior geometric capability in our partial scene registration task. It demonstrates that the pretraining objective of *generation* empowers the foundation models to have a decent capability of finding geometric correspondences in 3D scenes. Another observation is that video encoders generally perform better than image encoders. The reason is that video foundation models have a better understanding of object shapes and geometry within the scenes from the multi-view input frames.

## 4 Analysis

The purpose of this section is to provide additional exploration towards the optimal usage of visual foundation models. The selection of encoding methods requires consideration of the trade-off between memory usage, running time, and performance. We will dive into complexity analysis and the study of design choices for various and a combination of foundation models.





Figure 7: Evaluation on different video downsampling strategies for V-JEPA on the segmentation task. *Keyframe Sampling* samples every  $N$  frames to form a new video sequence, while *Clip Sampling* directly samples consecutive video clips. The performance before downsampling is regarded as 100%. Keyframe sampling demonstrates less performance drop with the same level of downsampling.

#### 4.1 Complexity Analysis

We compare memory usage, computation time, and model performance (*vision-language reasoning on ScanQA*) in Table 6 and Figure 6. Our findings show that image encoders generally require less time to process a sample compared to video and 3D encoders. And diffusion-based models, when used for feature extraction, require significantly more memory than other discriminative models. Noticeably, the drawbacks in running time become evident for 2D backbones, especially image encoders, when attempting to obtain a scene embedding by aggregating multi-view image embeddings. To illustrate this, we consider a 300-frame video as an exemplar of posed 2D information for a complex scene (a 10-second video at 30 FPS). As the length of the video increases, 2D methods, which necessitate feature extraction for each image frame, rapidly consume a substantial amount of time to process a single scene. In contrast, a 3D point encoder requires significantly less time to process a scene. Nevertheless, 3D encoders exhibit relatively poor model performance, which can be attributed to the scarcity of training data. To fully demonstrate their potential in scene understanding tasks, efforts should be directed toward enhancing the generalizability of 3D foundation models. All analyses and computations were conducted on an Nvidia A100 GPU.

#### 4.2 Ablation Study – Insights into Optimal Usage of Visual Foundation Models

**Video downsampling strategy.** Long and high frame-per-second videos take a lot of space to store and time to process. We explore two straightforward ways of conducting temporal downsampling to achieve more efficient processing without sacrificing too much performance. As shown in Figure 7, we explore the *keyframe sampling* (blue) and *clip sampling* (orange) strategies. We can observe that keyframe sampling is a better strategy than clip sampling in this setting, more wisely balancing the trade-off between video processing overhead and task performance.

**Combination of multiple encoders.** We explore whether a mixture of foundation models (experts) has the potential to strengthen the capability of 3D scene understanding. We experiment on the 3D semantic segmentation task with three feature sources: LSeg, StableDiffusion, and Swin3D. When combining different feature sources, we concatenate all features along the channel dimension for every point in the point cloud. The results are shown in Figure 8. After combining features from different sources, there exists potential that the semantic understanding capability can be boosted in a *mixture of experts* manner. However, it is not necessarily true that combining the best features will lead to the best performance. For example, LSeg (1) has stronger capability on semantic segmentation than StableDiffusion (2) and Swin3D (3) individually, but it is StableDiffusion + Swin3D (2+3) that reaches the best performance when combining two features together.

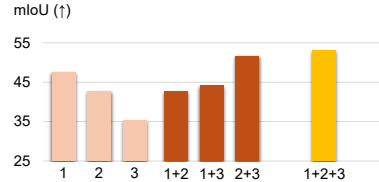


Figure 8: Evaluation on the segmentation task with (1) LSeg, (2) SD, (3) Swin3D, and their combinations.

**Appendix (supplementary material).** The appendix offers a comprehensive introduction of all of our evaluated models and additional experiment details, and includes more visualization and ablation experiments. We also elaborate on the limitations, broader impact, and future direction of our work.

## 5 Conclusion

This paper presents the first comprehensive analysis of leveraging visual foundation models for complex 3D scene understanding. We explore the strengths and weaknesses of models designed for various modalities and trained with different objectives. Our study reveals the superior performance of DINOv2, the advantages of video models in object-level tasks, and the benefits of diffusion models in geometric registration tasks. Surprisingly, we find limitations of language-pretrained models in language-related tasks. The extensive analysis suggests that a more flexible encoder selection can play a crucial role in future scene understanding and multi-modal reasoning tasks.

## References

- [1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1
- [2] P. Achlioptas, A. Abdelreheem, F. Xia, M. Elhoseiny, and L. Guibas. ReferIt3D: Neural listeners for fine-grained 3D object identification in real-world scenes. In *ECCV*, 2020. 6
- [3] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong. VATT: Transformers for multimodal self-supervised learning from raw video, audio and text. In *NeurIPS*, 2021. 3
- [4] M. Assran, Q. Duval, I. Misra, P. Bojanowski, P. Vincent, M. Rabbat, Y. LeCun, and N. Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, 2023. 3
- [5] D. Azuma, T. Miyanishi, S. Kurita, and M. Kawanabe. ScanQA: 3D question answering for spatial scene understanding. In *CVPR*, 2022. 1, 3, 5, 6, 17
- [6] M. E. Banani, A. Raj, K.-K. Maninis, A. Kar, Y. Li, M. Rubinstein, D. Sun, L. Guibas, J. Johnson, and V. Jampani. Probing the 3D awareness of visual foundation models. In *CVPR*, 2024. 1, 2, 3, 18
- [7] S. Banerjee and A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005. 6
- [8] H. Bao, L. Dong, S. Piao, and F. Wei. BeiT: Bert pre-training of image transformers. In *ICLR*, 2022. 3
- [9] D. Baranchuk, I. Rubachev, A. Voynov, V. Khrulkov, and A. Babenko. Label-efficient semantic segmentation with diffusion models. In *ICLR*, 2022. 3
- [10] A. Bardes, J. Ponce, and Y. LeCun. MC-JEPA: A joint-embedding predictive architecture for self-supervised learning of motion and content features. *arXiv preprint arXiv:2307.12698*, 2023. 3
- [11] A. Bardes, Q. Garrido, J. Ponce, X. Chen, M. Rabbat, Y. LeCun, M. Assran, and N. Ballas. V-JEPA: Latent video prediction for visual representation learning, 2024. URL <https://openreview.net/forum?id=WFYbBOE0tv>. 3, 5, 17
- [12] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, V. Jampani, and R. Rombach. Stable Video Diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3, 5, 16
- [13] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *NeurIPS*, 2020. 3
- [14] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 3
- [15] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3D: Learning from rgb-d data in indoor environments. In *3DV*, 2017. 4

- [16] D. Z. Chen, A. X. Chang, and M. Nießner. ScanRefer: 3D object localization in rgb-d scans using natural language. In *ECCV*, 2020. 6, 17
- [17] S. Chen, P. Sun, Y. Song, and P. Luo. DiffusionDet: Diffusion model for object detection. In *ICCV*, 2023. 3
- [18] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3
- [19] X. Chen and K. He. Exploring simple siamese representation learning. In *CVPR*, 2021. 3
- [20] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, 2017. 4, 5, 6, 7, 8, 17
- [21] R. Ding, J. Yang, C. Xue, W. Zhang, S. Bai, and X. Qi. PLA: Language-driven open-Vocabulary 3D scene understanding. In *CVPR*, 2023. 3, 4
- [22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3
- [23] Y. Duan, X. Guo, and Z. Zhu. DiffusionDepth: Diffusion denoising approach for monocular depth estimation. *arXiv preprint arXiv:2303.05021*, 2023. 3
- [24] C. Fang, X. Hu, K. Luo, and P. Tan. Ctrl-Room: Controllable text-to-3D room meshes generation with layout constraints. *arXiv preprint arXiv:2310.03602*, 2023. 1
- [25] R. Fridman, A. Abecasis, Y. Kasten, and T. Dekel. SceneScape: Text-driven consistent scene generation. In *NeurIPS*, 2023. 1
- [26] G. Gao, W. Liu, A. Chen, A. Geiger, and B. Schölkopf. GraphDreamer: Compositional 3D scene synthesis from scene graphs. In *CVPR*, 2024. 1
- [27] S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018. 3
- [28] R. Girdhar, A. El-Nouby, M. Singh, K. V. Alwala, A. Joulin, and I. Misra. OmniMAE: Single model masked pretraining on images and videos. In *CVPR*, 2023. 3
- [29] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent-a new approach to self-supervised learning. In *NeurIPS*, 2020. 3
- [30] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3
- [31] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 3
- [32] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3
- [33] L. Höllein, A. Cao, A. Owens, J. Johnson, and M. Nießner. Text2Room: Extracting textured 3D meshes from 2D text-to-image models. In *ICCV*, 2023. 1
- [34] Y. Hong, C. Lin, Y. Du, Z. Chen, J. B. Tenenbaum, and C. Gan. 3D concept learning and reasoning from multi-view images. In *CVPR*, 2023. 3, 4
- [35] Y. Hong, H. Zhen, P. Chen, S. Zheng, Y. Du, Z. Chen, and C. Gan. 3D-LLM: Injecting the 3D world into large language models. In *NeurIPS*, 2023. 1, 3, 4, 5, 6, 17
- [36] Y. Hong, Z. Zheng, P. Chen, Y. Wang, J. Li, and C. Gan. MultiPLY: A multisensory object-centric embodied large language model in 3D world. In *CVPR*, 2024. 1
- [37] J. Hou, B. Graham, M. Nießner, and S. Xie. Exploring data-efficient 3D scene understanding with contrastive scene contexts. In *CVPR*, 2021. 3, 5
- [38] A. Joulin, L. Van Der Maaten, A. Jabri, and N. Vasilache. Learning visual features from large weakly supervised data. In *ECCV*, 2016. 3
- [39] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, 1976. 17
- [40] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik. LERF: Language embedded radiance fields. In *ICCV*, 2023. 3

- [41] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 7, 17
- [42] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia. LISA: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 1
- [43] Y. LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1), 2022. 3
- [44] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022. 3, 5, 6, 16
- [45] J. Li, D. Li, C. Xiong, and S. Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 3, 6
- [46] J. Li, D. Li, S. Savarese, and S. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 3, 6, 17
- [47] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *ACL*, 2004. 6
- [48] H. Liu, M. Cai, and Y. J. Lee. Masked discrimination for self-supervised learning on point clouds. In *ECCV*, 2022. 3
- [49] J. Liu, G. Wang, Z. Liu, C. Jiang, M. Pollefeys, and H. Wang. RegFormer: An efficient projection-aware transformer network for large-scale point cloud registration. In *ICCV*, 2023. 7
- [50] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6, 17
- [51] C. Ma, Y. Yang, C. Ju, F. Zhang, J. Liu, Y. Wang, Y. Zhang, and Y. Wang. DiffusionSeg: Adapting diffusion towards unsupervised object discovery. *arXiv preprint arXiv:2303.09813*, 2023. 3
- [52] X. Ma, S. Yong, Z. Zheng, Q. Li, Y. Liang, S.-C. Zhu, and S. Huang. SQA3D: Situated question answering in 3D scenes. In *ICLR*, 2023. 1, 3, 5, 6, 17
- [53] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, 2018. 3
- [54] Y. Man, L.-Y. Gui, and Y.-X. Wang. BEV-Guided Multi-Modality Fusion for Driving Perception. In *CVPR*, 2023. 1
- [55] Y. Man, L.-Y. Gui, and Y.-X. Wang. Situational awareness matters in 3d vision language reasoning. In *CVPR*, 2024. 1, 3, 6
- [56] K. Namekata, A. Sabour, S. Fidler, and S. W. Kim. EmerDiff: Emerging pixel-level semantic knowledge in diffusion models. In *ICLR*, 2024. 3
- [57] S. Nasiriany, F. Xia, W. Yu, T. Xiao, J. Liang, I. Dasgupta, A. Xie, D. Driess, A. Wahid, Z. Xu, Q. Vuong, T. Zhang, T.-W. E. Lee, K.-H. Lee, P. Xu, S. Kirmani, Y. Zhu, A. Zeng, K. Hausman, N. Heess, C. Finn, S. Levine, and B. Ichter. PIVOT: Iterative visual prompting elicits actionable knowledge for VLMs. *arXiv preprint arXiv:2402.07872*, 2024. 1
- [58] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL <https://openreview.net/forum?id=a68SUt6zFt>. 2, 3, 5, 16
- [59] Y. Pang, W. Wang, F. E. Tay, W. Liu, Y. Tian, and L. Yuan. Masked autoencoders for point cloud self-supervised learning. In *ECCV*, 2022. 3
- [60] Z. Pang, Z. Xie, Y. Man, and Y.-X. Wang. Frozen transformers in language models are effective visual encoder layers. In *ICLR*, 2024. 3
- [61] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL*, 2002. 6
- [62] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 3
- [63] S. Peng, K. Genova, C. M. Jiang, A. Tagliasacchi, M. Pollefeys, and T. Funkhouser. Openscene: 3D scene understanding with open vocabularies. In *CVPR*, 2023. 3, 4

- [64] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, Q. Ye, and F. Wei. Kosmos-2: Grounding multimodal large language models to the world. In *ICLR*, 2024. 1
- [65] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3, 5, 6, 16
- [66] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(140):1–67, 2020. 3
- [67] M. Ranzinger, G. Heinrich, J. Kautz, and P. Molchanov. AM-RADIO: Agglomerative model–reduce all domains into one. In *CVPR*, 2024. 1, 3
- [68] H. Rasheed, M. Maaz, S. Shaji, A. Shaker, S. Khan, H. Cholakkal, R. M. Anwer, E. Xing, M.-H. Yang, and F. S. Khan. GLaMM: Pixel grounding large multimodal model. In *CVPR*, 2024. 1
- [69] M. Roberts, J. Ramapuram, A. Ranjan, A. Kumar, M. A. Bautista, N. Paczan, R. Webb, and J. M. Susskind. HyperSim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *ICCV*, 2021. 4
- [70] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3, 5, 16, 18
- [71] S. Saxena, A. Kar, M. Norouzi, and D. J. Fleet. Monocular depth estimation using diffusion models. *arXiv preprint arXiv:2302.14816*, 2023. 3
- [72] J. L. Schonberger and J.-M. Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 4
- [73] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev. Laion-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 16
- [74] J. Schult, S. Tsai, L. Höllein, B. Wu, J. Wang, C.-Y. Ma, K. Li, X. Wang, F. Wimbauer, Z. He, P. Zhang, B. Leibe, P. Vajda, and J. Hou. ControlRoom3D: Room generation using semantic proxy rooms. In *CVPR*, 2024. 1
- [75] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, P. Luo, A. Geiger, and H. Li. DriveLM: Driving with graph visual question answering. *arXiv preprint arXiv:2312.14150*, 2023. 1
- [76] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 3
- [77] L. Tang, M. Jia, Q. Wang, C. P. Phoo, and B. Hariharan. Emergent correspondence from image diffusion. In *NeurIPS*, 2023. 16
- [78] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1
- [79] X. Tian, J. Gu, B. Li, Y. Liu, C. Hu, Y. Wang, K. Zhan, P. Jia, X. Lang, and H. Zhao. DriveVLM: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024. 1
- [80] Z. Tong, Y. Song, J. Wang, and L. Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022. 3, 5
- [81] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *TPAMI*, 13(04):376–380, 1991. 17
- [82] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. CIDER: Consensus-based image description evaluation. In *CVPR*, 2015. 6
- [83] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, and Y. Qiao. VideoMAEv2: Scaling video masked autoencoders with dual masking. In *CVPR*, 2023. 3, 5
- [84] W. Wang, Z. Chen, X. Chen, J. Wu, X. Zhu, G. Zeng, P. Luo, T. Lu, J. Zhou, Y. Qiao, and J. Dai. VisionLLM: Large language model is also an open-ended decoder for vision-centric tasks. In *NeurIPS*, 2023. 1



- [85] Y. Wang, K. Li, Y. Li, Y. He, B. Huang, Z. Zhao, H. Zhang, J. Xu, Y. Liu, Z. Wang, S. Xing, G. Chen, J. Pan, J. Yu, Y. Wang, L. Wang, and Y. Qiao. InternVideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 3
- [86] W. Wu, Y. Zhao, M. Z. Shou, H. Zhou, and C. Shen. DiffuMNask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *ICCV*, 2023. 3
- [87] X. Wu, X. Wen, X. Liu, and H. Zhao. Masked scene contrast: A scalable framework for unsupervised 3D representation learning. In *CVPR*, 2023. 3
- [88] S. Xie, J. Gu, D. Guo, C. R. Qi, L. Guibas, and O. Litany. PointContrast: Unsupervised pre-training for 3D point cloud understanding. In *ECCV*, 2020. 3, 5
- [89] H. Xu, G. Ghosh, P.-Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, and C. Feichtenhofer. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In *EMNLP*, 2021. 3
- [90] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, 2023. 3
- [91] J. Xu, X. Zhou, S. Yan, X. Gu, A. Arnab, C. Sun, X. Wang, and C. Schmid. Pixel aligned language models. In *CVPR*, 2024. 1
- [92] S. Yan, Y. Yang, Y. Guo, H. Pan, P.-s. Wang, X. Tong, Y. Liu, and Q. Huang. 3D feature prediction for masked-autoencoder-based point cloud pretraining. In *ICLR*, 2024. 3
- [93] Y.-Q. Yang, Y.-X. Guo, J.-Y. Xiong, Y. Liu, H. Pan, P.-S. Wang, X. Tong, and B. Guo. Swin3D: A pretrained transformer backbone for 3D indoor scene understanding. *arXiv preprint arXiv:2304.06906*, 2023. 3, 5, 17
- [94] C. Yeshwanth, Y.-C. Liu, M. Nießner, and A. Dai. ScanNet++: A high-fidelity dataset of 3D indoor scenes. In *ICCV*, 2023. 4
- [95] Z. J. Yew and G. H. Lee. REGTR: End-to-end point cloud correspondences with transformers. In *CVPR*, 2022. 7, 8, 17
- [96] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, and J. Lu. Point-BERT: Pre-training 3D point cloud transformers with masked point modeling. In *CVPR*, 2022. 3
- [97] R. Zellers, J. Lu, X. Lu, Y. Yu, Y. Zhao, M. Salehi, A. Kusupati, J. Hessel, A. Farhadi, and Y. Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *CVPR*, 2022. 3
- [98] G. Zhan, C. Zheng, W. Xie, and A. Zisserman. What does stable diffusion know about the 3d scene? *arXiv preprint arXiv:2310.06836*, 2023. 3
- [99] J. Zhang, C. Herrmann, J. Hur, L. Polania Cabrera, V. Jampani, D. Sun, and M.-H. Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. In *NeurIPS*, 2023. 1, 3, 18
- [100] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *ECCV*, 2016. 3
- [101] R. Zhang, Z. Guo, P. Gao, R. Fang, B. Zhao, D. Wang, Y. Qiao, and H. Li. Point-M2AE: multi-scale masked autoencoders for hierarchical point cloud pre-training. In *NeurIPS*, 2022. 3
- [102] R. Zhang, J. Han, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, P. Gao, and Y. Qiao. LLaMA-Adapter: Efficient fine-tuning of language models with zero-init attention. In *ICLR*, 2024. 3
- [103] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. 3
- [104] Y. Zhang, Z. Gong, and A. X. Chang. Multi3DRefer: Grounding text description to multiple 3D objects. In *ICCV*, 2023. 6
- [105] Z. Zhang, R. Girdhar, A. Joulin, and I. Misra. Self-supervised pretraining of 3D features on any point-cloud. In *ICCV*, 2021. 3, 5
- [106] Z. Zhang, B. Yang, B. Wang, and B. Li. GrowSP: Unsupervised semantic segmentation of 3D point clouds. In *CVPR*, 2023. 7

- [107] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu. Unleashing text-to-image diffusion models for visual perception. In *ICCV*, 2023. 3
- [108] H. Zhen, X. Qiu, P. Chen, J. Yang, X. Yan, Y. Du, Y. Hong, and C. Gan. 3D-VLA: A 3D vision-language-action generative world model. In *ICML*, 2024. 1, 6
- [109] J. Zheng, J. Zhang, J. Li, R. Tang, S. Gao, and Z. Zhou. Structured3D: A large photo-realistic dataset for structured 3D modeling. In *ECCV*, 2020. 3, 4, 5
- [110] J. Zheng, J. Zhang, J. Li, R. Tang, S. Gao, and Z. Zhou. Structured3D: A large photo-realistic dataset for structured 3D modeling. In *ECCV*, 2020. 17
- [111] J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong. iBOT: Image bert pre-training with online tokenizer. In *ICLR*, 2022. 3
- [112] Y. Zhou, L. Huang, Q. Bu, J. Zeng, T. Li, H. Qiu, H. Zhu, M. Guo, Y. Qiao, and H. Li. Embodied understanding of driving scenarios. *arXiv preprint arXiv:2403.04593*, 2024. 1
- [113] Z. Zhu, X. Ma, Y. Chen, Z. Deng, S. Huang, and Q. Li. 3D-VisTA: Pre-trained transformer for 3D vision and text alignment. In *ICCV*, 2023. 3

# Lexicon3D: Probing Visual Foundation Models for Complex 3D Scene Understanding

Supplementary Material

## A Additional Experiment Details

In this section, we provide a detailed introduction of all the visual foundation models we have evaluated, including the checkpoints we use and how we extract feature representations from the encoder backbones. Our code will be publicly released for replication and further evaluation.

### A.1 Evaluated Visual Foundation Models

Our evaluation and analysis are conducted mainly on the seven models listed in Table 1 in the main body paper. We have chosen models such that they cover most of the backbones used by recent 3D scene understanding and reasoning work. In this part, we discuss all the models we have used in our experiments and explain their pretraining object, the dataset used for pretraining, the public checkpoints we choose, and the method we leverage to extract features from their backbones. We start with image foundation models, and then video and 3D models.

**DINOv2** [58]. DINOv2 leverages an image-wise contrastive objective by minimizing the distance of features from the same samples, and maximizing those from different samples. They also include a patch-wise denoising objective by performing reconstruction from masked inputs. They train their model on a large-scale image dataset, LVD-142M [58], which contains 142 million unlabeled images. We take the standard DINOv2 implementation<sup>1</sup> and use the pretrained ViT-L/14 checkpoint for our evaluations.

**LSeg** [44]. LSeg aims to align visual features from images with corresponding semantic information provided by natural language descriptions by maximizing the correlation between the text embedding and the image pixel embedding of the ground-truth class of the pixel. We use the official checkpoint<sup>2</sup> of ViT-L/16 that is trained on a mixture of seven datasets [44].

**CLIP** [65]. CLIP aligns visual and textual representations in a shared embedding space through contrastive learning by maximizing the similarity between the embeddings of corresponding image-caption pairs while minimizing the similarity of non-matching pairs. CLIP was trained on a large and diverse dataset of image-caption pairs sourced from the internet including over 400 million image-text pairs. We use the official implementation and checkpoint<sup>3</sup> with a ViT-L/14 as the backbone for our evaluations.

**StableDiffusion (SD)** [70]. SD is a diffusion-based model used for generating high-quality images from text prompts. The model is trained to gradually remove noise from images, transforming random noise into coherent images that match the provided text descriptions. This model is trained on LAION5B [73] which contains over five billion of images paired with detailed captions. We follow DIFT [77]<sup>4</sup> to extract features from SD and we use the checkpoint SD2.1 for our evaluation. We use the features from block index 1 for all tasks. The noise timestep is set to 100 by default. We use null-prompt as the text condition.

**StableVideoDiffusion (SVD)** [12]. SVD is an extension of SD from image generation to video generation by incorporating additional temporal modules. SVD is first initialized from an image-level pretrained diffusion checkpoint (SD2.1), then is further finetuned on 10 million videos. We use their publicly released image-to-video variant (SVD-xt)<sup>5</sup>. We build our feature extractor pipeline following DIFT [77] and extract the features from index 1 for all tasks. The noise timestep is set to 25 by default. We use the first-frame image as the condition for all the cross-attention modules while we use the unconditional version for the latent input of the UNet – we concatenate an all-zero vector

---

<sup>1</sup><https://github.com/facebookresearch/dinov2>

<sup>2</sup><https://github.com/is1-org/lang-seg>

<sup>3</sup><https://github.com/openai/CLIP>

<sup>4</sup><https://github.com/Tsingularity/dift>

<sup>5</sup><https://huggingface.co/stabilityai/stable-video-diffusion-img2vid-xt>

to the framewise embeddings. Each time, we feed a video clip of 25 frames to SVD to process the features.

**V-JEPA** [11]. V-JEPA aims to learn robust visual representations by predicting future states of visual data. This model is pretrained on a mixed video dataset containing over 2 million of videos. We take their official implementation <sup>6</sup> and the checkpoint ViT-L/16 under resolution of  $224 \times 224$ . We obtain their per-patch representation by removing the last pooling and linear layers. Each time, we feed a video clip of 16 frames to V-JEPA to process the features.

**Swin3D** [93]. Swin3D adapts the Swin Transformer to handle 3D data, such as point clouds and volumetric data. We use the official checkpoint <sup>7</sup> that takes Swin3D-L as the backbone and is pretrained on the Structure3D dataset [110] with semantic segmentation as the target.

## A.2 Additional Evaluation Details for Vision-Language Scene Reasoning

**Datasets.** We evaluate the performance on two challenging indoor 3D VQA datasets: ScanQA [5] and SQA3D [52]. SQA3D features over 33K QA pairs, while ScanQA consists of more than 41K pairs. Each entry in these datasets includes a complex 3D indoor scene, a question, and corresponding answers. We use the splits provided by the respective datasets.

**Optimization.** We keep the LLM parameters frozen and finetune the shallow visual projection Q-Former module [46] to align features from different encoders to the LLM input space. Different from [35], we train the Q-Former module from scratch for a fair comparison of all encoders. Following the approach of 3D-LLM, we pretrain the module for 10 epochs using 3D-Language dataset [35] and then finetune it on the training split of the two evaluation datasets for 35 epochs. Both stages use the AdamW [50] optimizer with a linear warmup and cosine decay learning rate scheduler. While longer training can further improve performance, trends stabilize after 35 training epochs.

## A.3 Additional Evaluation Details for Registration

**Dataset generation.** When generating corresponding partial scene point clouds from ScanNet dataset, due to memory constraint, we downsample the partial scene point clouds to 4,096 points each with the farthest point sampling (FPS) algorithm, if the number of points in  $P_1$  and  $P_2$  is over 4,096. We follow the same train/val split on the semantic segmentation task in our partial scene registration task.

**Optimization.** We follow REGTR [95] to adopt a Transformer cross-encoder module to enable cross-reasoning of the foundation model features from two point clouds, followed by a lightweight decoder to obtain the corresponding position of each point in the other point cloud, forming altogether  $N_1 + N_2$  pairs of correspondences, where  $N_1$  and  $N_2$  are the number of points in  $P_1$  and  $P_2$ , respectively. Afterward, the rotation  $\mathbf{R}$  and the translation  $\mathbf{t}$  can be obtained in a closed-form solution solved by a weighted version of the Kabsch-Umeyama [39, 81] algorithm. We use Adam [41] for optimization and train our model for 30 epochs.

## A.4 License of Dataset Used

In this section, we list the licenses of all the datasets we have used during our evaluation:

- ScanNet [20]: MIT License.
- ScanQA [5]: CC BY-NC-SA 3.0 License.
- SQA3D [52]: CC-BY-4.0 License.
- ScanRefer [16]: CC BY-NC-SA 3.0 License.
- 3DLanguage-Data [35]: MIT License.

In addition, we utilize a number of public foundation model checkpoints pretrained on various data sources in our paper. Please refer to their original paper for the license of datasets they have used in pretraining their models.

<sup>6</sup><https://github.com/facebookresearch/jepa>

<sup>7</sup><https://github.com/microsoft/Swin3D>

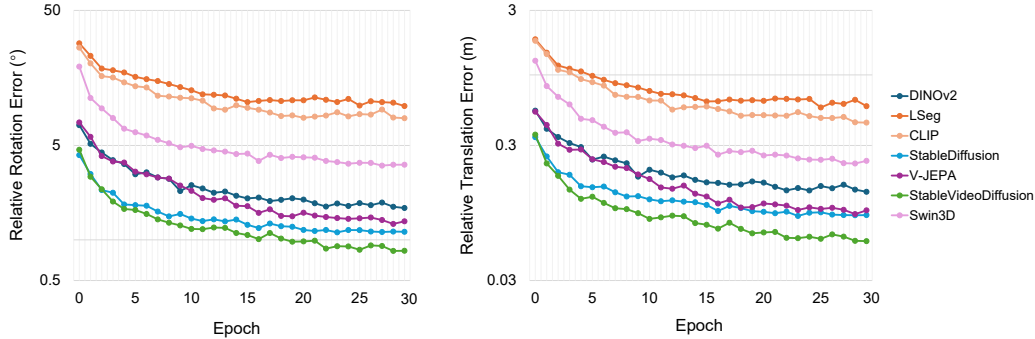


Figure A: Evaluation curves of Relative Rotation Error (RRE) and Relative Translation Error (RTE) on the partial scene registration task during different training stages.

Stable Diffusion	BLEU-1	BLEU-4	METEOR	ROUGE	CIDEr
<i>Evaluation of noise level <math>t</math></i>					
$t = 1 \text{ step}$	35.3	11.6	14.0	34.5	68.5
$t = 25 \text{ steps}$	35.6	11.5	14.0	34.2	68.3
<b><math>t = 100 \text{ steps}</math></b>	35.5	11.7	14.1	34.9	68.2
$t = 200 \text{ steps}$	34.3	10.9	13.9	33.9	66.6
<i>Evaluation of feature layer <math>l</math></i>					
$l = 0$	33.6	10.5	13.3	32.6	65.9
<b><math>l = 1</math></b>	35.5	11.7	14.1	34.9	68.2
$l = 2$	34.9	11.4	14.0	34.5	68.0

Table 7: Evaluation of diffusion noise level and feature layers when using StableDiffusion [70] for feature extraction. The setting we choose are highlighted in **bold**.

## B Additional Experiment Results

### B.1 Evaluation Curves during Different Training Stages

We show the evaluation curves for the partial scene registration in Figure A. We can observe that the performance ranking of different foundation models stays mainly unchanged throughout the training process.

### B.2 Diffusion Noise Level and Feature Layer

In Table 7, we evaluate the effect of different noise level (*noise steps*) and different feature layers in the decoder module in leveraging StableDiffusion (SD) [70] for feature extraction. The results show that for SD, adding noise  $t < 100$  steps in general leads to the best performance. When  $t$  increases beyond 100 steps, the performance starts to downgrade. As for decoder layers, the decoding portion of the UNet consists of 4 blocks. We skip the final layer closest to the output and consider layers 0, 1, and 2. The results demonstrate that the output features of the layer one decoder lead to the best performance. These observations are consistent with the study in [6, 99].

### B.3 Additional Qualitative Results

We show additional qualitative results for partial scene registration in Figure B, demonstrating that the family of StableDiffusion and StableVideoDiffusion which use the objective of generative pretraining obtains superior performance. In addition, video encoders like V-JEPA and StableVideoDiffusion are equipped with a stronger capability to find geometric correspondences.



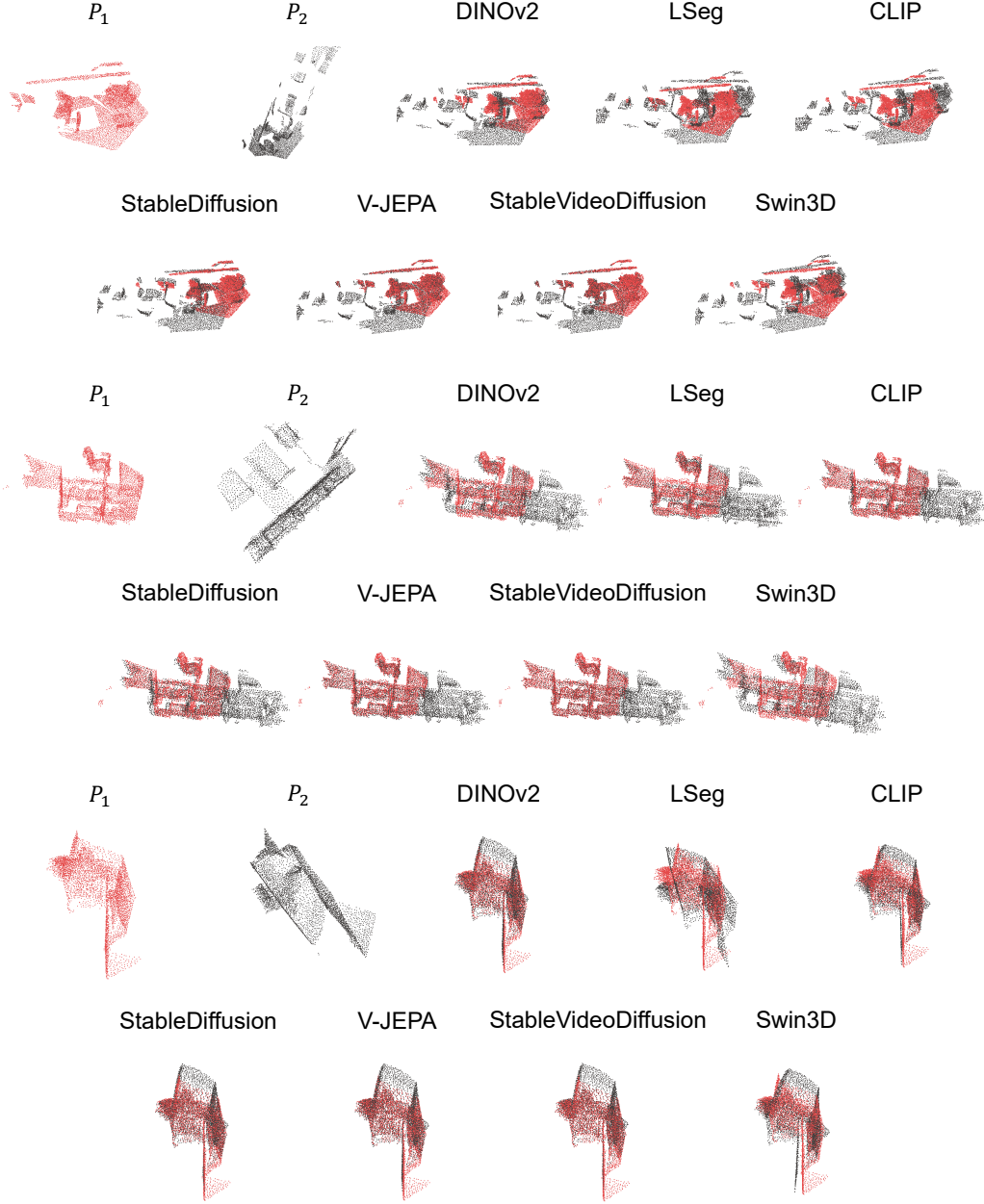


Figure B: Visualization of partial scene registration results. The family of generative models StableDiffusion and StableVideoDiffusion obtains superior performance. Also, video encoders like V-JEPA and StableVideoDiffusion have better geometric understanding capability than image encoders.

## C Limitations and Future Work

Although we have made a substantial effort to explore the role of visual foundation models in various scene understanding tasks, our perception of this problem remains relatively limited. This section provides a detailed discussion of the limitations and outlines potential future directions.

**Model capacities not strictly identical or comparable.** Our evaluation focuses on seven vision foundation models due to their availability and common use in recent work. Consequently, all our experiments are based on publicly available checkpoints released by the project owners. Although we have attempted to choose models with similar capacities, achieving strictly identical backbone archi-

textures was not possible without re-training all the baselines ourselves. However, such experiments require an enormous amount of computational resources that we cannot afford.

**Our evaluation focuses on indoor scenarios.** Recent literature often separates the study of indoor scene perception and reasoning from outdoor scenarios, which are often relevant to autonomous driving or robotics applications. Outdoor scenarios present different challenges compared to indoor scenes. Lexicon3D focuses its evaluation solely on indoor scenes. While this is a valid choice considering that most scene-level multi-modal benchmarks are still based on indoor scenes, it is not comprehensive. Outdoor scenarios contain large ego-movement speeds and many more dynamic moving objects than indoor scenes. Evaluating these scenes will likely lead to unique observations, and we consider this a direct future direction.

**We adopt the most straightforward approach to probing.** To evaluate the capabilities of the visual foundation models, we freeze their parameters and only tune the linear or shallow probing head. This approach allows us to analyze the pretrained methods’ capabilities without altering their models through the finetuning process. While we argue that probing the frozen encoder provides the most accurate understanding of these models, we acknowledge that the ability to quickly adapt to new tasks with finetuning is also an important aspect of an encoder. However, finetuning these large-scale models, which often have close to billion-level parameters, requires a significant amount of time and computational resources. We leave this study for future work.

## D Societal Impact

We anticipate a potential positive social impact from our work. Lexicon3D represents one of the first steps towards a comprehensive understanding of large-scale visual foundation models in real-world 3D scene analysis and reasoning. This understanding could lead to the development of more robust and efficient scene encoding systems, benefiting a wide range of applications, including autonomous driving, virtual reality, household robots, and multi-modal chatbots. Ultimately, this could contribute to a more inclusive, efficient, and safer world, where technology understands and adapts to the diverse ways humans perceive and navigate their environments.

**Potential negative societal impact.** We do not see a direct negative societal impact on our work. Indirect potential negative impact involves misusing strong scene encoding foundation models for surveillance or virtual reality. We believe it is crucial for researchers to proactively consider these concerns and establish guidelines to ensure the responsible usage of these models.