3D-VL-FLY: A MULTI-TASK 3D VISION–LANGUAGE BENCHMARK FOR AERIAL POST-DISASTER DAMAGE ASSESSMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in multimodal large language models have enabled sophisticated visual understanding and reasoning for disaster response. Emerging benchmarks such as DisasterM3 have begun to extend multimodal learning to post-disaster scenarios through tasks such as visual question answering and damage assessment. However, these benchmarks predominantly rely on 2D imagery, limiting the evaluation of multimodal models' ability to reason over the three-dimensional geometric structure of disaster environments. Important evidence, including collapsed roof structures, debris extent, building-level spatial relationships, and partially occluded structures, are difficult to fully characterize from a single 2D view.

In this study, we introduce 3D-VL-Fly, the first 3D vision–language benchmark specifically designed for post-disaster damage assessment. Built from 3D point-cloud scenes generated from aerial imagery collected following Hurricane Ian (2022), the benchmark provides building-level instance segmentation in addition to semantic segmentation of roads and trees, attributes per building independently, and question–answer pairs spanning eight question families with spatially grounded building references. Beyond visual question answering, 3D-VL-Fly incorporates both *referring segmentation*, which evaluates a model's ability to localize language-referred buildings or regions, and *reasoning segmentation*, which evaluates whether a model can identify the 3D evidence relevant to answering a reasoning-oriented question. This unified formulation enables the evaluation of multimodal models not only on what they answer, but also on *where* the relevant evidence is located in a 3D disaster scene. We evaluate a range of pretrained 3D and 2D vision–language models and encoders, and fine-tune a representative 3D vision–language model. Released models remain below a simple majority baseline on VQA and achieve low segmentation accuracy, indicating substantial room for improvement.

1 INTRODUCTION

Natural disasters can cause widespread damage to the built environment, making rapid and accurate damage assessment critical for coordinating emergency response, allocating resources, and planning recovery (Chen et al., 2026; Wang et al., 2026; Goudemant & Francesconi, 2026; Chen et al., 2025a; Li et al., 2026; Amini et al., 2025; Xiao et al., 2026). Aerial imagery has become an increasingly important data source in this setting because it enables rapid coverage of affected areas that may be difficult or dangerous for human teams to access. In particular, unmanned aerial vehicles (UAVs) have attracted increasing attention from disaster response teams because they can be deployed rapidly after an event to capture high-resolution imagery across large affected areas (Zhang et al., 2026; Sarkar & Rahnemoonfar, 2023; Rahnemoonfar et al., 2021). However, the large volume of imagery collected after disasters makes manual interpretation time-consuming and highlights the need for automated methods that can efficiently interpret complex disaster scenes and support downstream decision-making.

The availability of aerial imagery has facilitated the development of task-specific datasets and deep learning methods for a range of disaster-response tasks, including building damage assessment (Gupta et al., 2019; Manzini et al., 2024; Chen et al., 2025a; Adriano et al., 2021; Lee et al.,

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

2022), change detection and captioning (Chen et al., 2025b), semantic and instance segmentation (Rahnemoonfar et al., 2023; 2021; Tran et al., 2025), and visual question answering (Sarkar et al., 2023; Sarkar & Rahnemoonfar, 2023; Zhang et al., 2026). While these task-specific approaches have enabled substantial progress in automated disaster-scene understanding, they typically address individual tasks in isolation. More recently, multimodal large language models (MLLMs) have enabled the development of multi-modal, multi-task foundation models for remote-sensing applications, supporting diverse capabilities such as visual question answering, image and region captioning, visual grounding, and segmentation (Kuckreja et al., 2024; Irvin et al., 2025; Xin et al., 2026; Zhang et al., 2024; Zhan et al., 2025; Quenum et al., 2025). Despite advances in remote-sensing models for processing aerial imagery and geospatial data, post-disaster analysis remains constrained by the limited availability of comprehensive, domain-specific datasets that support diverse multimodal tasks while capturing the fine-grained spatial and semantic complexity of disaster scenes.

Recent efforts such as DisasterM3 (Wang et al., 2026) have introduced multimodal tasks including visual question answering (VQA) and damage assessment. Together, aforementioned benchmarks have advanced the evaluation of vision and vision–language models in post-disaster scenarios. Despite this progress, existing disaster benchmarks predominantly rely on 2D imagery. This limits their ability to evaluate whether multimodal models can reason over the three-dimensional geometric structure of disaster environments (Taciroglu et al., 2025a;b; Le et al., 2025). Such evidence may be ambiguous or entirely invisible from a single 2D view, whether nadir or oblique (Tsujimoto et al., 2026; Zhu et al., 2025b); and some of the evidence that matters for post-disaster assessment is inherently three-dimensional: building height, collapsed or deformed roof structures, the spatial extent and height of debris, and structures partially occluded by trees or other buildings. As a result, a model that performs well on 2D benchmarks may still lack the geometric understanding required for reliable assessment in real-world settings.

Recent work such as 3D-AeroRelief (Le et al., 2025) has begun to incorporate 3D representations into post-disaster analysis, but primarily focuses on semantic segmentation. This motivates extending beyond 3D perception toward multimodal models that can jointly understand 3D geometry, language, and damage-related evidence.

In parallel, the 3D vision–language community has developed benchmarks for question answering, visual grounding, and spatial reasoning (Azuma et al., 2022; Chen et al., 2020; Achlioptas et al., 2020; Ma et al., 2023), as well as 3D-aware language models (Huang et al., 2024a; 2023). More recently, LISA-3D has extended language-guided segmentation to 3D through multi-view geometric consistency (Guo et al., 2025). These efforts demonstrate the potential of combining language with 3D visual representations, but have focused primarily on indoor or general-purpose scenes rather than outdoor disaster environments. In particular, damage-specific 3D spatial reasoning and language-grounded localization remain underexplored.

To address these gaps, we introduce 3D-VL-Fly, to the best of our knowledge, the first 3D vision–language benchmark specifically designed for post-disaster damage assessment. 3D-VL-Fly is built from 3D point clouds reconstructed from UAV aerial imagery collected following Hurricane Ian (2022) in Florida. The benchmark provides building-level instance segmentation together with semantic segmentation of roads and trees, independent attribute annotations for each building, and question–answer pairs spanning eight question families with spatially grounded building references. Beyond VQA, 3D-VL-Fly incorporates two segmentation-based tasks: *referring segmentation*, which evaluates a model's ability to localize buildings or regions described in natural language, and *reasoning segmentation* (Lai et al., 2023), which evaluates whether a model can identify the 3D evidence relevant to answering a reasoning-oriented question. This unified formulation enables the evaluation of multimodal models not only on *what* they answer, but also on *where* the relevant evidence is located in a 3D disaster scene. We evaluate representative 2D and 3D vision–language models on 3D-VL-Fly across VQA, language-guided segmentation, and building-attribute recognition. Our results show that released 3D vision–language models remain well below a training-majority reference on VQA, and achieve low mIoU on language-guided segmentation, highlighting substantial room for improvement in 3D multimodal reasoning for disaster scenes.

In summary, our contributions are as follows:

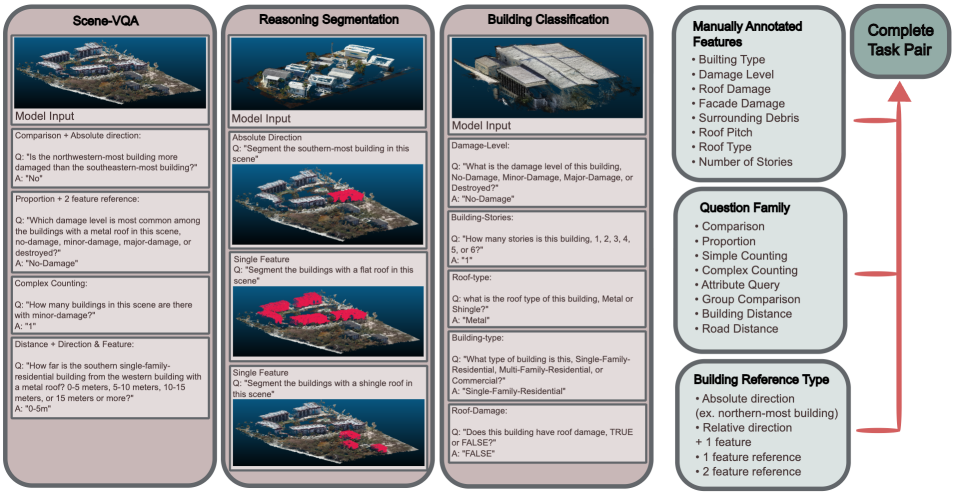


Figure 1: 3D-VL-Fly at a glance. Given a 3D point cloud of a hurricane-affected scene reconstructed from UAV imagery, 3D-VL-Fly evaluates multimodal models on three complementary tasks: **(a)** answering scene-level questions about building damage, counts, attributes, and distances; **(b)** segmenting the buildings referred to by directional or attribute-based descriptions; and **(c)** classifying fine-grained damage and structural attributes of individual buildings. Because the same building references are shared across tasks, the benchmark measures not only *what* a model answers but also *where* the relevant evidence lies in 3D.

- We introduce 3D-VL-Fly, the first 3D vision–language benchmark for post-disaster damage assessment, built from point clouds of a real hurricane-affected area.
- We provide rich, multi-level annotations, including building instance segmentation, semantic segmentation of roads and trees, per-building damage and structural attributes, and spatially grounded question–answer pairs across wight question categories.
- We benchmark 17 pretrained models and encoders across VQA, language-guided segmentation, building-attribute recognition, and building instance segmentation, together with a fine-tuning experiment, and analyze their strengths and limitations to establish baselines for future research.

2 Related Work

2.1 Vision–Language Models

General-purpose models. The potential of vision–language foundation models was unleashed by CLIP (Radford et al., 2021), whose large-scale contrastive image–text pretraining enabled strong zero-shot transfer. BLIP (Li et al., 2022) extended this paradigm to generative modeling, unifying image–text alignment with text generation. Flamingo (Alayrac et al., 2022) and BLIP-2 (Li et al., 2023) subsequently bridged pretrained vision encoders with frozen large language models (LLMs), enabling richer interaction between visual and textual representations. Building on these architectures, visual instruction tuning (Liu et al., 2023b; Dai et al., 2023) substantially improved reasoning and open-ended dialogue capabilities, and has since become a cornerstone of state-of-the-art multimodal large language models (MLLMs), including the Llama (Grattafiori et al., 2024) and Qwen-VL (Bai et al., 2025) families. Nevertheless, because these models generate outputs as discrete tokens, localization, when supported, is typically serialized as textual coordinates of bounding boxes or points (Bai et al., 2023; Chen et al., 2023; Peng et al., 2023), following the coordinates-as-tokens formulation of Pix2Seq (Chen et al., 2022). Such sparse representations cannot capture the precise object boundaries required by dense prediction tasks such as referring segmentation (Hu et al., 2016). To overcome this limitation, LISA (Lai et al., 2023) augmented an MLLM with a dedicated mask decoder, using a special segmentation token whose embedding is decoded into a pixel-level mask. Beyond enabling referring segmentation, LISA introduced *reasoning segmentation*, in which the model must infer the target region from an implicit query that requires world knowledge

or reasoning rather than an explicit description. Subsequent works such as PixelLM (Ren et al., 2024), LLM-Seg (Wang & Ke, 2024), and GLaMM (Rasheed et al., 2024) extended this paradigm to multi-target segmentation or grounded conversation generation.

A parallel trajectory has unfolded in 3D. Early unified 3D vision–language transformers (Zhu et al., 2023) learned joint representations of point clouds and text, and were followed by 3D-aware LLMs that ingest point clouds or object-centric scene representations to support 3D question answering, captioning, and dialogue (Hong et al., 2023; Huang et al., 2024a; 2023; Chen et al., 2024a; Xu et al., 2024). Mirroring the 2D progression from textual outputs to dense masks, language-guided segmentation has likewise been extended to 3D, including reasoning segmentation over point clouds (Huang et al., 2025; He et al., 2024) and multi-view geometrically consistent formulations such as LISA-3D (Guo et al., 2025). More recent generalist models unify these capabilities within a single framework: Grounded 3D-LLM (Chen et al., 2024b) links language to scene regions through dedicated referent tokens, and 3D-LLaVA (Deng et al., 2025) couples an LLM with a unified superpoint-based encoder that supports both 3D question answering and referring segmentation. However, these models have been trained and evaluated almost exclusively on indoor scenes, and their ability to reason over and localize damage-related evidence in outdoor disaster environments remains unexplored.

Remote sensing and post-disaster models. In remote sensing, domain-adapted contrastive models such as RemoteCLIP (Liu et al., 2024) improved image–text alignment for overhead imagery. Following the success of general-purpose MLLMs, a range of remote-sensing MLLMs have been proposed, including RSGPT (Hu et al., 2023), GeoChat (Kuckreja et al., 2024), EarthGPT (Zhang et al., 2024), and SkyEyeGPT (Zhan et al., 2025), which support VQA, image and region captioning, and visual grounding within a single model, while TEOChat (Irvin et al., 2025) extends this paradigm to temporal sequences of Earth observation imagery. Mirroring the general-domain shift toward dense outputs, pixel-level language grounding has been explored through referring remote-sensing image segmentation (Yuan et al., 2024) and MLLM-based reasoning and grounded segmentation (Quenum et al., 2025; Shabbir et al., 2025; Xin et al., 2026). In disaster response, models have remained largely task-specific: building damage assessment has been addressed with dedicated change-detection and segmentation networks (Zheng et al., 2021; Adriano et al., 2021), and only recently have general-purpose and remote-sensing MLLMs begun to be evaluated and adapted on disaster imagery (Wang et al., 2026; Zhang et al., 2026). All of these remote-sensing and disaster models, however, operate on 2D imagery, and the ability of 3D vision–language models to reason over and localize damage-related evidence in outdoor disaster environments remains unexplored.

2.2 VISION–LANGUAGE BENCHMARKS.

General-purpose benchmarks. Vision–language evaluation began with image captioning (Young et al., 2014; Chen et al., 2015) and later expanded to visual question answering (Antol et al., 2015; Goyal et al., 2017) and compositional reasoning (Hudson & Manning, 2019). In parallel, the RefCOCO family (Kazemzadeh et al., 2014; Yu et al., 2016; Mao et al., 2016) and gRefCOCO (Liu et al., 2023a) established standard benchmarks for referring segmentation. As MLLMs acquired pixel-level outputs, new benchmarks were introduced to evaluate them: ReasonSeg (Lai et al., 2023) targets reasoning segmentation with implicit queries, MUSE (Ren et al., 2024) and LLM-Seg40K (Wang & Ke, 2024) extend reasoning segmentation to multi-target and large-scale settings, and GranD (Rasheed et al., 2024) evaluates grounded conversation generation, in which textual responses are interleaved with segmentation masks. These benchmarks have since been extended to remote sensing by SegEarth-R2 (Xin et al., 2026), which spans multi-granularity, multi-target, and implicit reasoning queries, and GeoSeg (Jiang et al., 2026), which provides a diagnostic benchmark with hierarchical difficulty levels. However, all of these benchmarks are built on 2D images and therefore cannot evaluate whether models reason over explicit 3D geometry or localize evidence in 3D space.

The 3D vision–language community has developed analogous benchmarks on reconstructed indoor scenes, primarily derived from ScanNet (Dai et al., 2017). ScanRefer (Chen et al., 2020), ReferIt3D (Achlioptas et al., 2020), and Multi3DRefer (Zhang et al., 2023) evaluate language-based object grounding in point clouds; Scan2Cap (Chen et al., 2021) addresses dense 3D captioning; and ScanQA (Azuma et al., 2022) and SQA3D (Ma et al., 2023) evaluate 3D question answering and situated reasoning. These benchmarks, however, are confined to room-scale indoor scenes contain-

ing household objects. Outdoor 3D datasets such as SemanticKITTI (Behley et al., 2019), Waymo Open (Sun et al., 2020), nuScenes (Caesar et al., 2020), and KITTI-360 (Liao et al., 2022) extend 3D understanding to large-scale urban environments, but capture sparse, street-level observations from vehicle-mounted sensors. Aerial datasets such as DALES (Varney et al., 2020), SensatUrban (Hu et al., 2022), and FRACTAL (Gaydon et al., 2024) offer substantially broader spatial coverage from an overhead perspective. Yet these outdoor datasets target perception tasks such as semantic segmentation and detection, and provide no language annotations for evaluating 3D vision–language reasoning.

Post-disaster benchmarks. Aerial and satellite imagery has driven the development of numerous task-specific datasets for disaster response. For building damage assessment, xBD (Gupta et al., 2019) established a large-scale benchmark for pre- and post-event satellite imagery, followed by multimodal and multi-sensor datasets (Adriano et al., 2021; Chen et al., 2025a) and UAV-based building damage datasets (Lee et al., 2022; Manzini et al., 2024). UAV imagery has also enabled high-resolution semantic segmentation of post-disaster scenes, including FloodNet (Rahnemoonfar et al., 2021) and RescueNet (Rahnemoonfar et al., 2023), as well as datasets for wildfire-affected land cover (Tran et al., 2025) and disaster change captioning (Chen et al., 2025b). Visual question answering was introduced to disaster analysis through RescueNet-VQA (Sarkar & Rahnemoonfar, 2023) and SAM-VQA (Sarkar et al., 2023), which pose questions about damage, flooding, and scene conditions in UAV imagery. More recently, DisasterBench (Zhang et al., 2026) and DisasterM3 (Wang et al., 2026) have scaled this direction to multi-task, multi-sensor evaluation of MLLMs, covering VQA, damage assessment, and related reasoning tasks across diverse disaster events. These benchmarks represent an important step toward general-purpose disaster-response assistants, yet they remain limited to 2D imagery and do not evaluate reasoning over the 3D geometry of disaster scenes.

A small number of efforts have begun to exploit 3D representations of disaster-affected areas. Following the 2025 Los Angeles wildfires, UAV-based reconnaissance campaigns released 3D point clouds, orthomosaics, and digital surface models of the Eaton and Palisades burn areas (Taciroglu et al., 2025a;b), though without semantic or damage annotations. 3D-AeroRelief (Le et al., 2025) introduced the first annotated point-cloud benchmark for post-disaster semantic segmentation. These works demonstrate the value of 3D data for damage analysis, but they are limited to raw reconnaissance data or perception tasks and do not incorporate language. To the best of our knowledge, no benchmark jointly evaluates question answering, referring segmentation, and reasoning segmentation over 3D disaster scenes. 3D-VL-Fly addresses this gap by bringing the 3D vision–language paradigm, previously confined to indoor or intact urban scenes, to real post-disaster environments with building-level, spatially grounded annotations.

3 Dataset and Methodology

3D-VL-Fly is built from point clouds reconstructed from UAV imagery collected after Hurricane Ian (2022) in Florida. The dataset contains 64 scenes with 308 annotated buildings. Building on the point clouds from 3D-AeroRelief (Le et al., 2025), in this study, we add building instance masks, per-building damage and structural attributes, geographic North references, and language queries. In this section, we first describe the supported tasks, followed by the annotation process for building instances and attributes and the generation of language queries. Quantitative experiments are conducted on a 49-scene subset, as described in Section 4.1.

Task Definitions and Evaluation Scope 3D-VL-Fly supports four tasks. (i) Scene-level VQA: given a scene point cloud and a natural-language question, predict the answer. (ii) Language-guided segmentation: given a scene point cloud and a natural-language reference to one or more buildings, predict a point-level mask of the referenced building(s). This task comprises referring segmentation and reasoning segmentation. (iii) Building attribute recognition: given a cropped point cloud of a single building and a target attribute, predict the attribute class. (iv) Building instance segmentation: partition all buildings in a scene into individual instances without a language query. The manually annotated instance masks serve as ground truth for all four tasks.



Figure 2: Example image pairs annotators used to assign class labels.

Scene and Building Segmentation The 3D-AeroRelief point clouds (Le et al., 2025) provide semantic labels for buildings, trees, roads, and background, but do not distinguish individual buildings. Using CloudCompare (CloudCompare development team, 2025), we isolated the building-labeled points in each scene and manually partitioned them into individual structures, assigning each a unique instance ID. This yields 308 building instances across 64 scenes, which serve as reference masks for training and evaluation.

Building Annotations To provide detailed and consistent annotations for each building across all scenes, we adopted a hybrid 2D/3D annotation methodology. Annotators were provided with high-resolution 2D images of the buildings alongside renderings of their 3D counterparts. Because building IDs are inherently tied to the 3D spatial data, we projected the 3D scenes onto 2D image planes based on specific camera perspectives. We then padded the 3D mask of each building to capture its immediate surroundings and isolated the corresponding regions in the 2D images. This extraction yielded multiple 2D crops for each building, significantly streamlining the annotation process (see Figure 2).

Our feature selection focused on visible building damage and characteristics relevant to post-disaster assessment. We aimed to strike a practical balance regarding annotation granularity: seeking more detail than the simple binary damaged/undamaged classification used in existing datasets, while avoiding overly specific feature classes that would be difficult to reliably annotate from our reference imagery. Each building is annotated with its *building type* (single-family residential, multi-family residential, commercial, or unknown), *roof damage* and *facade damage* (true/false), *surrounding debris* (none, small amount, or large amount), *roof type* (metal, shingle, or other), *roof pitch* (low, high, or flat), *number of stories* (1–6), and *damage level*. Following xBD (Gupta et al., 2019), damage level is graded on four classes.

Each building was independently reviewed by three separate annotators. We then applied an aggregation to determine the majority label for each feature class. When all three annotators assigned different classes, we excluded that feature label for the building. Such cases accounted for less than 1% of the annotated samples.

Scene-level Question-answer Design For scene-level VQA, models receive a full 3D scene point cloud alongside a natural language question and, when necessary, a spatial grounding reference (e.g., the global direction of North). Answering a grounded query requires identifying the referenced building or buildings and evaluating the requested property. We structured our questions into eight distinct categories: comparison, proportion, simple counting, complex counting, attribute recognition, group comparison, building distance, and road distance. Additionally, building references are divided into three types: absolute direction, relative (light) direction combined with a feature, and identification via two distinct features.

North-reference estimation. We obtained a signed geographic-north direction from the source drone videos using synchronized flight telemetry and reviewed image-to-map registrations. Frame inventories, contact sheets, and available edit-timeline metadata were used to relate edited exports to their original flights. Camera cuts, incompatible camera models, and unsupported tracking intervals were treated as boundaries rather than bridged by interpolation. Local video mosaics were used to identify candidate neighborhoods in north-up NOAA Hurricane Ian imagery, not as orthophotos. Automatic matches based on image features, roofs, roads, water, and building layouts supplied candidate registrations; acceptance required visual inspection of corresponding physical features and rendered overlays (see section A).

Language-Guided Segmentation The language-guided segmentation task evaluates selection of buildings using spatial or attribute constraints. In this task, models are prompted with a natural

Table 1: Released-model VQA results (%). ND denotes non-directional questions ($n = 684$); D denotes directional questions ($n = 332$). Acc. is normalized exact match; Macro is the mean over represented question families. The same text-only majority reference applies to both input groups.

Model	ND Acc.	ND Macro	D Acc.	D Macro
Native point-cloud input				
3D-LLaVA (Deng et al., 2025)	23.39	21.00	20.48	19.56
Grounded 3D-LLM (Chen et al., 2024b)	1.46	2.37	–	–
Rendered-scene image input				
Qwen2.5-VL-7B (Bai et al., 2025)	37.28	29.76	13.55	12.50
InternVL3-8B (Zhu et al., 2025a)	23.68	23.76	13.86	12.81
LLaVA-OneVision-7B (Li et al., 2024)	29.53	30.71	14.16	13.06
Training-majority reference	44.88	41.61	45.78	45.37

language reference to a specific building (or group of buildings) using absolute direction, relative direction combined with a feature, a single unique feature, or a combination of two features. The model is then required to output a point-level segmentation mask corresponding to the referenced structure(s). Absolute direction, relative direction plus feature, and two feature references are constructed to uniquely identify a single building. Single feature references may instead correspond to one or multiple buildings sharing the specified attribute.

Furthermore, we provide an isolated building feature classification dataset to train models specifically on fine-grained attribute recognition. For this task, the model receives a cropped 3D point cloud of a single building along with its immediate surroundings (using the same 2-meter padding applied during the annotation phase). The objective is to correctly identify the designated feature class from this localized building point cloud, without the broader context of the full scene.

4 Results

4.1 Evaluation Protocol

We evaluate 17 pretrained models and encoders on the 49 scenes of Areas 3–8 (220 buildings). Areas 3–6 provide 28 training and five validation scenes, and Areas 7–8 provide 16 held-out test scenes with 85 buildings, all of which are evaluated. The test split contains 684 non-directional and 332 directional VQA questions and 584 and 331 segmentation prompts, respectively. Directional inputs additionally receive the North vector in the point-cloud frame. Because the two subsets contain different questions, they are not a controlled ablation of North conditioning. VQA and segmentation use released checkpoints without fine-tuning. Point-cloud models receive geometry with natural RGB, and image–language models receive three orthographic renderings with metric scene dimensions. VQA is scored by normalized exact match, with macro accuracy averaged over question families and compared against a per-family training-majority reference. Segmentation is scored by mIoU against the union of referenced building masks. Attribute recognition probes frozen image and point-cloud encoders with per-attribute logistic-regression classifiers trained on training buildings and tuned on validation, over 658 test records (Appendix B).

4.2 Illustrative Scene Questions

Figure 1 shows samples from dataset questions from 3 different training scenes. Counting queries require a scene-wide inventory; directional attribute queries additionally require resolving a building reference using the supplied North vector. These examples illustrate the task definitions and reference answers. They are not selected model successes or failures, and cannot explain the aggregate scores on their own.

4.3 Benchmark Results

Table 1 summarizes VQA performance. Within the rendered-scene comparison, Qwen2.5-VL has the highest non-directional exact-match accuracy (37.28%), while LLaVA-OneVision has the highest family-macro accuracy (30.71%). Among the native point-cloud models, 3D-LLaVA leads VQA

Table 2: Released-model language-guided segmentation results (% mIoU). Targets are whole-building unions at original point resolution. Non-directional and directional subsets contain 584 and 331 prompts, respectively.

Model	Non-directional	Directional
3D-LLaVA	1.41	1.58
Ground3D-LMM (released)	2.26	8.54
Grounded 3D-LLM	3.15	1.81
Reason3D (Huang et al., 2025)	8.34	5.30

Table 3: Building-attribute results (%) on 658 held-out attribute records. Encoders are frozen; attribute classifiers are fitted on training buildings and selected on validation. Macro averages the eight attribute accuracies. These are not zero-shot scene-model results.

Encoder	Accuracy	Attribute macro
Paired photograph/render composites		
DINOv2-Base (Oquab et al., 2023)	74.01	74.09
SigLIP2-Base (Tschannen et al., 2025)	73.10	73.38
OpenCLIP ViT-B/32 (Ilharco et al., 2021)	72.19	72.54
ConvNeXt-Base (Liu et al., 2022)	67.63	68.06
ResNet-50 (He et al., 2016)	67.33	67.74
Training-majority reference	53.95	54.56
Padded point-cloud crops		
OpenShape (Liu et al., 2023c)	58.21	58.81
Sonata (Wu et al., 2025)	40.73	41.86
Point-MAE (Pang et al., 2022)	36.78	37.40
Training-majority reference	53.95	54.56

on both subsets, at 23.39% and 20.48% accuracy. All evaluated VQA models remain below the corresponding training-majority reference. These figures measure short-answer agreement under the stated protocol; verbose or improperly formatted answers can fail exact match even when part of the response is relevant. A zero score does not distinguish incorrect scene interpretation from answer-format mismatch or generation failure; the aggregate results do not isolate these causes.

Table 2 reports mask overlap. Reason3D records the highest released-model segmentation mIoU on non-directional prompts (8.34%); Ground3D-LMM leads on directional prompts (8.54%), followed by Reason3D (5.30%). The low overlap scores indicate that reliable selection and delineation of the requested buildings remain unresolved under these transfer protocols. They do not, by themselves, identify dataset size or insufficient training as the cause; these models were not trained on the benchmark dataset.

In Table 3, DINOv2 yields the highest overall image-attribute accuracy (74.01%), followed by SigLIP2 (73.10%). OpenShape leads the point-cloud attribute comparison at 58.21%, above the 53.95% majority reference. Point-MAE and Sonata remain below that reference. These results concern attribute prediction from supplied building regions, not detection or unrestricted scene reasoning. Per-attribute results are reported in table 6 because class imbalance and variation in attribute difficulty are obscured by a single aggregate. For example, DINOv2 reaches 98.80% on building type and 100.00% on stories, while damage-level accuracy is 63.53% (61.18% for the majority reference) and roof-damage accuracy is 67.06% (71.76% for the reference). Aggregate attribute accuracy therefore should not be read as damage-assessment accuracy.

As a supplemental, class-agnostic building-instance test, OpenIns3D (Huang et al., 2024b) obtains AP_{25} and AP_{50} of 0.00% over the 16 test scenes. No building instances survive its fixed proposal and class-lookup thresholds. This is a failure of the evaluated configuration to produce accepted instances, rather than evidence that every possible proposal has zero overlap. This result is kept separate from language-guided segmentation, which has different inputs and target masks.

4.4 Separate Fine-Tuning Experiment

To examine adaptation to the dataset, we also report the earlier joint VQA and segmentation fine-tuning experiment with the 4B Ground3D-LMM configuration. Training covered all 2,087 eligible

Table 4: Earlier Ground3D-LMM adaptation experiment on the non-directional test subset (%). The released row is the archived evaluation accompanying that experiment, not the later released-model rerun in Table 2.

Configuration	VQA Acc.	VQA Macro	Seg. mIoU
Ground3D-LMM released (earlier evaluation)	0.00	0.00	2.39
Ground3D-LMM fine-tuned (epoch 4)	41.52	36.46	17.32
Training-majority reference	44.88	41.61	—

non-directional training records per epoch: 1,076 VQA records and 1,011 segmentation records. The run lasted five epochs; the epoch-four checkpoint was selected by validation VQA macro accuracy. Language adaptation used rank-32 LoRA updates, with the point encoder, cross-modal bridge, and mask decoder also trainable. This was not a LoRA-only update of an otherwise frozen system.

The jointly fine-tuned configuration reaches 41.52% VQA accuracy and 17.32% segmentation mIoU, compared with 0.00% and 2.39%, respectively, for its archived released-model evaluation. VQA accuracy and macro accuracy nevertheless remain below the training-majority reference. The result provides evidence of adaptation on the selected split, but does not establish robust directional reasoning or consistent gains across training seeds. The fine-tuned model was not evaluated on the directional subset.

Reporting limits. These are task-specific reference measurements, not a single ranking across input modalities. Attribute heads are trained; the released scene models are not. The test scenes come from two held-out areas, and the experiments use one fixed seed rather than repeated independent training runs. Scene-cluster bootstrap intervals for the released scene models are reported in Appendix B. They quantify variation among the available test scenes, not uncertainty across additional sites or training seeds. Appendix B gives additional preprocessing and adaptation settings.

5 Limitations

The dataset covers 64 scenes from a single hurricane, and testing is restricted to 16 scenes from two held-out areas. Results therefore do not establish generalization to other disasters, sensors, or reconstruction conditions. The reported scores are task-specific reference measurements rather than a single ranking across input modalities: attribute probes are trained, whereas the released scene models are not. No repeated-seed or directional fine-tuning study is available. Annotators used high-resolution 2D imagery alongside 3D renderings, which the 3D-only models do not receive, so some fine visual distinctions may be less observable in the model inputs. Automatically generated questions may contain recurring answer patterns, which the strong majority reference makes important to control for. Exact match may also reject semantically correct answers. Damage labels describe visible evidence and are not engineering assessments of structural safety.

6 Conclusion

We introduced 3D-VL-Fly, to the best of our knowledge the first 3D vision–language benchmark for post-disaster damage assessment. Built from UAV-derived point clouds of areas affected by Hurricane Ian, it provides building instance masks, per-building damage and structural attributes, geographic North references, and spatially grounded language queries. These support scene-level VQA, language-guided segmentation, building-attribute recognition, and building instance segmentation. Our evaluation of 17 pretrained models and encoders shows that current systems transfer poorly to this setting. No released model reaches a training-majority reference on VQA, image–language models degrade sharply on directional questions, and language-guided segmentation stays below 9% mIoU. Frozen point-cloud encoders also lag image encoders on attribute recognition, and damage-related attributes remain close to the majority reference even for the strongest image encoders. Fine-tuning on 3D-VL-Fly substantially improves both VQA and segmentation, which suggests that the benchmark supplies useful supervision, but a clear gap remains. We hope 3D-VL-Fly will support the development of 3D multimodal models that reason over geometry, language, and damage evidence together. Promising directions include models that combine 3D geometry with high-resolution imagery, robust use of geographic orientation, and extension of the benchmark to additional disaster events and regions.

AI USE STATEMENT

In this work, we used generative AI tools for paraphrasing, grammar correction, improving the clarity and readability of the manuscript, and retrieval and discovery to find related work. We have not used generative AI tools for generating research ideas, designing experiments, conducting analyses, interpreting results, or generating scientific claims, and other uses of generative AI are not applicable to this work. Additionally, we used generative AI tools for language editing and enhancement of the manuscript. We reviewed all AI-assisted text and verified that the final wording accurately reflects the authors' intended meaning. All scientific content, claims, interpretations, and conclusions were independently reviewed by the authors. We take responsibility for the final content of this work, including any text produced or modified with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, the 3D-VL-Fly dataset, including point clouds, annotations, and question–answer pairs, will be publicly released upon publication. Additional details on dataset construction, annotation, model implementation, and evaluation are provided in the Appendix.

REFERENCES

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European Conference on Computer Vision*, pp. 422–440, 2020.

Bruno Adriano, Naoto Yokoya, Junshi Xia, Hiroyuki Miura, Wen Liu, Masashi Matsuoka, and Shunichi Koshimura. Learning from multimodal and multitemporal earth observation data for building damage mapping. *ISPRS Journal of Photogrammetry and Remote Sensing*, 175:132–143, 2021.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Kooshan Amini, Yuhao Liu, Jamie Ellen Padgett, Guha Balakrishnan, and Ashok Veeraraghavan. Debris segmentation using post-hurricane aerial imagery. *Computer-Aided Civil and Infrastructure Engineering*, 40(25):4116–4131, 2025.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

Daichi Azuma, Taiki Miyanishi, Shintaro Kurita, and Motoi Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19129–19139, 2022.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *Proc. of the IEEE/CVF International Conf. on Computer Vision (ICCV)*, 2019.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European Conference on Computer Vision*, pp. 202–221, 2020.

Hongruixuan Chen, Jian Song, Olivier Dietrich, Clifford Broni-Bediako, Weihao Xuan, Junjue Wang, Xinlei Shao, Yimin Wei, Junshi Xia, Cuiling Lan, et al. Bright: A globally distributed multimodal building damage assessment dataset with very-high-resolution for all-weather disaster response. *Earth System Science Data*, 17(11):6217–6253, 2025a.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal LLM's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. In *International Conference on Learning Representations (ICLR)*, 2022.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Runsen Xu, Ruiyuan Lyu, Dahua Lin, and Jiangmiao Pang. Grounded 3D-LLM with referent tokens. *arXiv preprint arXiv:2405.10370*, 2024b.

Zhaohui Chen, Elyas Asadi Shamsabadi, Sheng Jiang, Luming Shen, and Daniel Dias-da Costa. Integration of large vision language models for efficient post-disaster damage assessment and reporting. *Nature Communications*, 17(1):1481, 2026.

Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X. Chang. Scan2Cap: Context-aware dense captioning in RGB-D scans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Zhenyuan Chen, Chenxi Wang, Ningyu Zhang, and Feng Zhang. Rscc: A large-scale remote sensing change caption dataset for disaster events. *arXiv preprint arXiv:2509.01907*, 2025b.

CloudCompare development team. Cloudcompare (version 2.13.2) [gpl software], 2025. `https://www.cloudcompare.org/`.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jiajun Deng, Tianyu He, Li Jiang, Tianyu Wang, Feras Dayoub, and Ian Reid. 3D-LLaVA: Towards generalist 3D LMMs with omni superpoint transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Charles Gaydon, Michel Daab, and Floryne Roche. Fractal: An ultra-large-scale aerial lidar dataset for 3d semantic segmentation of diverse landscapes, 2024. URL `https://arxiv.org/abs/2405.04634`.

Thomas Goudemant and Benjamin Francesconi. Optimizing latent representations for robust building damage assessment onboard earth observation satellites. *arXiv preprint arXiv:2605.29575*, 2026.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Zhongbin Guo, Jiahe Liu, Wenyu Gao, Yushan Li, Chengzhi Li, and Ping Jian. Lisa-3d: Lifting language-image segmentation to 3d via multi-view consistency. *arXiv preprint arXiv:2512.01008*, 2025.

Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. xBD: A dataset for assessing building damage from satellite imagery. *arXiv preprint arXiv:1911.09296*, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Shuting He, Henghui Ding, Xudong Jiang, and Bihan Wen. SegPoint: Segment any point cloud via large language model. In *European Conference on Computer Vision (ECCV)*, 2024.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3D-LLM: Injecting the 3D world into large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Qingyong Hu, Bo Yang, Sheikh Khalid, Wen Xiao, Niki Trigoni, and Andrew Markham. Sensaturban: Learning semantics from urban-scale photogrammetric point clouds. *International Journal of Computer Vision*, 130(2):316–343, 2022.

Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *European Conference on Computer Vision (ECCV)*, pp. 108–124, 2016.

Yuan Hu, Jianlong Yuan, Congcong Wen, Xiaonan Lu, and Xiang Li. RSGPT: A remote sensing vision language model and benchmark. *arXiv preprint arXiv:2307.15266*, 2023.

Haifeng Huang, Zehan Wang, Rongjie Huang, Luping Liu, Xize Cheng, Yang Zhao, Tao Jin, and Zhou Zhao. Chat-3d v2: Bridging 3d scene and large language models with object identifiers. *arXiv preprint arXiv:2312.08168*, 2023.

Kuan-Chih Huang, Xiangtai Li, Lu Qi, Shuicheng Yan, and Ming-Hsuan Yang. Reason3D: Searching and reasoning 3D segmentation via large language model. In *International Conference on 3D Vision (3DV)*, 2025.

Yining Huang et al. Leo: An embodied generalist agent in 3d world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a.

Zhening Huang, Xiaoyang Wu, Xi Chen, Hengshuang Zhao, Lei Zhu, and Joan Lasenby. OpenIns3D: Snap and lookup for 3D open-vocabulary instance segmentation. In *European Conference on Computer Vision*, pp. 169–185, 2024b. doi: 10.1007/978-3-031-73033-7_10. URL https://arxiv.org/abs/2309.00616.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP. Zenodo, 2021. URL https://doi.org/10.5281/zenodo.5143773.

Jeremy Andrew Irvin, Emily Ruoyu Liu, Joyce Chuyi Chen, Ines Dormoy, Jinyoung Kim, Samar Khanna, Zhuo Zheng, and Stefano Ermon. Teochat: A large vision-language assistant for temporal earth observation data. In *Proceedings of the International Conference on Learning Representations*, 2025.

Lifan Jiang, Yuhang Pei, Boxi Wu, Yan Zhao, Tianrun Wu, Shulong Yu, Lihui Zhang, and Deng Cai. GeoSeg: Training-free reasoning-driven segmentation in remote sensing imagery. *arXiv preprint arXiv:2603.03983*, 2026.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27831–27840, 2024.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model, arxiv. *arXiv preprint arXiv:2308.00692*, 3, 2023.

Nhut Le, Ehsan Karimi, and Maryam Rahnemoonfar. 3daerorelief: The first 3d benchmark uav dataset for post-disaster assessment. *arXiv preprint arXiv:2509.11097*, 2025.

Cheng-Chun Lee, Navjot Kaur, Ali Mahdavi-Amiri, and Ali Mostafavi. Ida-bd: pre- and post-disaster high-resolution satellite imagery for building damage assessment from hurricane ida, 2022. URL https://www.designsafe-ci.org/data/browser/public/designsafe.storage.published/PRJ-3563.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. URL https://arxiv.org/abs/2408.03326.

Hao Li, Fabian Deuser, Wenping Yin, Steffen Knoblauch, Wufan Zhao, Filip Biljecki, Yong Xue, and Wei Huang. Towards generative location awareness for disaster response: A probabilistic cross-view geolocalization approach. *ISPRS Journal of Photogrammetry and Remote Sensing*, 237:130–145, 2026.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. In *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 2022.

Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23592–23601, 2023a.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. RemoteCLIP: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023b.

Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. OpenShape: Scaling up 3D shape representation towards open-world understanding. *arXiv preprint arXiv:2305.10764*, 2023c. URL https://arxiv.org/abs/2305.10764.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022. URL https://arxiv.org/abs/2201.03545.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *International Conference on Learning Representations*, 2023.

Thomas Manzini, Priyankari Perali, Raisa Karnik, and Robin Murphy. Crasar-u-droids: A large scale benchmark dataset for building alignment and damage assessment in georectified suas imagery. *arXiv preprint arXiv:2407.17673*, 2024.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11–20, 2016.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL https://arxiv.org/abs/2304.07193.

Yatian Pang, Wenxiao Wang, Francis E. H. Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *European Conference on Computer Vision*, pp. 604–621, 2022. doi: 10.1007/978-3-031-20086-1_35. URL https://arxiv.org/abs/2203.06604.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

Jerome Quenum, Wen-Han Hsieh, Tsung-Han Wu, Ritwik Gupta, Trevor Darrell, and David M Chan. Lisat: Language-instructed segmentation assistant for satellite imagery. *arXiv preprint arXiv:2505.02829*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Roberson Murphy. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654, 2021.

Maryam Rahnemoonfar, Tashnim Chowdhury, and Robin Murphy. Rescuenet: A high resolution uav semantic segmentation dataset for natural disaster damage assessment. *Scientific data*, 10(1): 913, 2023.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13009–13018, 2024.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. PixelLM: Pixel reasoning with large multimodal model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26374–26383, 2024.

Argho Sarkar and Maryam Rahnemoonfar. Rescuenet-vqa: A large-scale visual question answering benchmark for damage assessment. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1150–1153. IEEE, 2023.

Argho Sarkar, Tashnim Chowdhury, Robin Roberson Murphy, Aryya Gangopadhyay, and Maryam Rahnemoonfar. Sam-vqa: Supervised attention-based visual question answering model for post-disaster damage assessment on remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–16, 2023.

Akashah Shabbir, Mohammed Zumri, Mohammed Bennamoun, Fahad S. Khan, and Salman Khan. GeoPixel: Pixel grounding large multimodal model in remote sensing. In *International Conference on Machine Learning (ICML)*, 2025.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Ertugrul Taciroglu, Mohammad Askari, Jeffrey Berman, Mustafa Cetinkaya, Karen Dedinsky, Devasmit Dutta, Negar Elhami Khorasani, Melis Fidansoy, Erica Fischer, Elias Gueidon, Michael Grilliot, Patrick Hadinata, Julia Hutchins, Debasish Jana, Ercan Kaya, Matias Korfmacher, Yiwen Liang, Andrew Lyda, Sriram Narasimhan, Serdar Selamet, Riyaaz Shaik, Joseph Wartman, Maxwell Wondolowski, Zafer YILMAZ, and Jaqueline Zdebski. Palisades fire burn area, in infrastructure resilience and impact study for wildfire-affected infrastructure in los angeles, 2025a. URL https://www.designsafe-ci.org/data/browser/public/designsafe.storage.published/PRJ-5815/#detail-29a05904-6e55-4089-abbc-411e881586a0/?version=3.

Ertugrul Taciroglu, Mohammad Askari, Jeffrey Berman, Mustafa Cetinkaya, Karen Dedinsky, Devasmit Dutta, Negar Elhami Khorasani, Melis Fidansoy, Erica Fischer, Elias Gueidon, Michael Grilliot, Patrick Hadinata, Julia Hutchins, Debasish Jana, Ercan Kaya, Matias Korfmacher, Yiwen Liang, Andrew Lyda, Sriram Narasimhan, Serdar Selamet, Riyaaz Shaik, Joseph Wartman, Maxwell Wondolowski, Zafer YILMAZ, and Jaqueline Zdebski. Eaton fire burn area, in infrastructure resilience and impact study for wildfire-affected infrastructure in los angeles, 2025b. URL https://www.designsafe-ci.org/data/browser/public/designsafe.storage.published/PRJ-5815/#detail-5a77ce86-fe4a-4394-b0c4-d12362e3885d/?version=3.

Anh Tran, Minh Tran, Esteban Marti, Jackson Cothren, Chase Rainwater, Sandra Eksioglu, and Ngan Le. Land8fire: A complete study on wildfire segmentation through comprehensive review, human-annotated multispectral dataset, and extensive benchmarking. *Remote Sensing*, 17(16): 2776, 2025.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. URL https://arxiv.org/abs/2502.14786.

Mai Tsujimoto, Junjue Wang, Weihao Xuan, and Naoto Yokoya. Geo3dvqa: Evaluating vision-language models for 3d geospatial reasoning from aerial imagery. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4994–5003. IEEE, 2026.

Nina Varney, Vijayan K Asari, and Quinn Graehling. Dales: A large-scale aerial lidar data set for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 186–187, 2020.

Junchi Wang and Lei Ke. Llm-seg: Bridging image segmentation and large language model reasoning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1765–1774. IEEE, 2024.

Junjue Wang, Weihao Xuan, Heli Qi, Zhihao Liu, Kunyi Liu, Yuhan Wu, Hongruixuan Chen, Jian Song, Junshi Xia, Zhuo Zheng, et al. Disasterm3: A remote sensing vision-language dataset for disaster damage assessment and response. *Advances in Neural Information Processing Systems*, 38, 2026.

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 *CVPR*, 2025.

Yiming Xiao, Archit Gupta, Miguel Esparza, Yu-Hsuan Ho, Antonia Sebastian, Hannah Weas, Rose Houck, and Ali Mostafavi. Recov-vision: Linking street view imagery and vision-language models for post-disaster recovery. *International Journal of Disaster Risk Reduction*, pp. 106388, 2026.

Zepeng Xin, Kaiyu Li, Luodi Chen, Wanchen Li, Xiao Yuchen, Hui Qiao, Weizhan Zhang, Deyu Meng, and Xiangyong Cao. Segearth-r2: Towards comprehensive language-guided segmentation for remote sensing images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13199–13210, 2026.

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 (ECCV)*, 2024.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *European Conference on Computer Vision (ECCV)*, 2016.

Zhenghang Yuan, Lichao Mou, Yuansheng Hua, and Xiao Xiang Zhu. RRSIS: Referring remote sensing image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 2024.

Yang Zhan, Zhitong Xiong, and Yuan Yuan. Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 221:64–77, 2025.

Tan Zhang, Quanyou Li, Lu Zhang, Jun Liu, Xiaofeng Zhu, and Ping Hu. Disasterbench: A multimodal benchmark for uav-based disaster response in complex environments. *arXiv preprint arXiv:2606.06217*, 2026.

Wei Zhang, Miaoxin Cai, Tong Zhang, Yin Zhuang, and Xuerui Mao. Earthgpt: A universal multimodal large language model for multisensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2024.

Yiming Zhang, ZeMing Gong, and Angel X. Chang. Multi3DRefer: Grounding text description to multiple 3D objects. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Zhuo Zheng, Yanfei Zhong, Junjue Wang, Ailong Ma, and Liangpei Zhang. Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters. *Remote Sensing of Environment*, 265:112636, 2021.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025a. URL https://arxiv.org/abs/2504.10479.

Xiao Xiang Zhu, Sining Chen, Fahong Zhang, Yilei Shi, and Yuanyuan Wang. Globalbuildingatlas: an open global and complete dataset of building polygons, heights and lod1 3d models. *Earth System Science Data*, 17(12):6647–6668, 2025b.

Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3D-VisTA: Pre-trained transformer for 3D vision and text alignment. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

A NORTH-REFERENCE ESTIMATION

We obtain a signed geographic-North direction from the source drone videos using synchronized flight telemetry and manually reviewed image-to-map registrations. Frame inventories, contact sheets, and available edit-timeline metadata are used to relate edited video exports to their original flights. Camera cuts, incompatible camera models, and unsupported tracking intervals are treated as boundaries rather than bridged by interpolation. Local video mosaics are used only to locate candidate neighborhoods in north-up NOAA Hurricane Ian imagery, not as orthophotos. Automatic matching of image features, roofs, roads, water, and building layouts provides candidate registrations, which are accepted only after visual inspection of corresponding physical features and rendered overlays.

Absolute direction anchors. Trusted frames supplied the absolute orientation for each connected tracking component. Compatible DJI or Parrot telemetry provided camera attitude after synchronization with source-frame time. Alternatively, manually reviewed correspondences on a common plane, such as a roof, pool, or road surface, defined an image-to-map homography. For an oriented homography H_a mapping undistorted anchor-frame pixels to a north-up map, and camera intrinsic matrix K_a, the camera-coordinate north direction is

$$\mathbf{n}_a = \mathrm{unit}\Big(K_a^{-1} H_a^{-1} \begin{bmatrix} 0 & -1 & 0 \end{bmatrix}^{\mathsf{T}}\Big), \tag{1}$$

where $\mathrm{unit}(\mathbf{v}) = \mathbf{v}/\|\mathbf{v}\|_2$. The map direction is upward in image coordinates, and its sign is checked against the reviewed alignment. All directions are stored in OpenCV camera coordinates: x points right, y points down, and z points forward. Camera intrinsics, available lens-distortion coefficients, calibration provenance, and the applicable image domain are recorded for each source. The map correspondence and camera calibration must use the same image domain.

Propagation through camera rotation. The implementation tracks image features with pyramidal Lucas–Kanade optical flow and a forward–backward consistency check, falling back to RootSIFT matching when fewer than 80 correspondences remain. Correspondences are undistorted before robust homography fitting with USAC_MAGSAC, or RANSAC when the former is unavailable. Adjacent-frame fits require at least 60 inliers; longer-baseline fits require at least 45. Both require an inlier ratio of at least 0.25 and an inlier convex hull covering at least 3.5% of each image. The respective median reprojection-error limits are 1.5 and 2.0 pixels in the resized tracking images. Calibrated homography decomposition supplies candidate rotations; the implementation selects the smallest-angle rotation and applies the source-specific rotation limit. Translation and plane-distance estimates are not used. Adjacent-frame rotations form a camera-rotation track, while longer-baseline estimates provide additional correction and disagreement diagnostics.

Let R_t map directions from a component's reference camera to frame t. An anchor direction is expressed in that reference frame as $\mathbf{m}_a = R_a^{\mathsf{T}} \mathbf{n}_a$. Between consecutive anchors, the implementation interpolates their reference-frame directions on the unit sphere and transforms the result into the current camera frame:

$$\mathbf{n}_t = R_t \ \mathrm{slerp}\left(\mathbf{m}_a, \mathbf{m}_b; \alpha_t\right), \tag{2}$$

where α_t is the relative frame-index position between anchors a and b. Before the first anchor and after the last anchor within a connected component, the nearest anchor's reference-frame direction is retained and propagated by the rotation track. Components without an absolute anchor remain unresolved. Where configured for direct publication, a reviewed direction measurement takes precedence over the propagated estimate. Check frames are compared with the pose-only prediction before that substitution. The checking scripts also support comparisons with separately reviewed map evidence, synchronized telemetry, and stable reconstruction rotations; reconstructions with orientation discontinuities are treated as diagnostics rather than independent confirmation.

Local projection and use with building instances. For an oblique or distorted image, the apparent direction of north varies across the frame. To use the output with a building mask, the direction is evaluated at that building's image centroid rather than copied from an image-center arrow. Let $\mathbf{q}_t(\mathbf{p})$ be the undistorted camera ray through image location $\mathbf{p}$, scaled to have third coordinate one, and let π_t denote projection with the stored intrinsics and distortion model. The implemented local direction is

$$\mathbf{d}_t(\mathbf{p}) = \mathrm{unit}[\pi_t(\mathbf{q}_t(\mathbf{p}) + \epsilon \mathbf{n}_t) - \pi_t(\mathbf{q}_t(\mathbf{p}))], \qquad \epsilon = 10^{-4}. \tag{3}$$

This normalized finite difference gives a signed image-plane tangent without estimating a building's depth or ground scale. The outputs contain per-frame directions, camera-model metadata, source identifiers, and tracking status in CSV and JSON form. They are orientation records, not building footprints or building-level absolute and relative direction labels. Using a camera-coordinate direction with a 3D scene additionally requires the registered scene-to-camera rotation Q_t, giving $\mathbf{n}_{\text{scene}} = Q_t^{\mathsf{T}}\mathbf{n}_t$. This coordinate conversion and the building-reference selection rules are distinct from the north-estimation procedure; the orientation pipeline does not determine metric distances or geographic building locations.

In addition to directional metadata, we automatically computed pairwise building distances and the distance from each building to the nearest road. To prevent the generation of road-related queries in scenes lacking significant infrastructure, we compiled a subset of scenes where road points constitute at least 10% of the point cloud. By combining the aggregated annotations, directional references, and computed distances, we generated scene-level VQA pairs across the eight question families.

B Additional Benchmark Details

We retain the per-attribute breakdown and scene-cluster uncertainty to distinguish aggregate performance from attribute-specific and scene-specific variation.

Model interfaces. Scene inputs originate from a fixed cache of up to 128,000 natural-RGB points per scene. Native models retain their respective sampling, voxelization, normalization, and generation settings; they do not share an identical point budget. Ground3D-LMM, Reason3D, and Grounded 3D-LLM use geometry-derived superpoint mappings. 3D-LLaVA uses deterministic 1-m geometric grouping and nearest-point mask reprojection. LL3DA uses 40,000 sampled points with RGB, normals, and height. The rendered-scene comparison uses three 768×768 orthographic views and deterministic decoding with at most 64 new tokens.

Reason3D uses its released task checkpoint with the SPFormer backbone permitted by the model's released documentation; byte identity with the original training-time backbone file is not established. Grounded 3D-LLM uses its released step-three checkpoint, public ScanNet class text, and its official grounding-task cue. These are explicitly adapted outdoor transfer protocols, not reproductions of the models' original indoor benchmark results.

Attribute features. OpenShape uses 10,000 sampled XYZRGB points, with zero-mean, unit-radius XYZ normalization. Point-MAE uses 1,024 XYZ points and concatenated mean/max pooling of unmasked encoder tokens; it does not consume RGB. Sonata uses reconstructed point features, geometry-derived normals, and mean pooling. Image encoders use their respective pretrained image preprocessing. Multi-view image features are averaged before classification. The sole missing paired-image building removes eight training labels; all 85 test buildings have paired views. Consequently, the 2D attribute split has 854/112/658 training/validation/test records, versus 862/112/658 for 3D.

Adaptation settings. The separate Ground3D run uses AdamW with learning rate 10^{-4}, weight decay 0.05, eight-microbatch gradient accumulation, and a cosine learning-rate schedule with warm-up. LoRA rank is 32, scaling is 64, and dropout is 0.1. The recorded trainable-parameter counts are 23,592,960 LoRA parameters, 10,960,384 point-encoder parameters, 1,447,936 bridge parameters, and 6,404,352 mask-decoder parameters. Five epochs cover all 2,087 eligible training records once per epoch; validation selects epoch four. No directional fine-tuning result is claimed.

Fine-tuning experiment. The joint VQA and segmentation run uses the 4B Ground3D-LMM configuration and covers all 2,087 eligible non-directional training records per epoch (1,076 VQA and 1,011 segmentation). Language adaptation uses rank-32 LoRA, with the point encoder, cross-modal bridge, and mask decoder also trainable. The epoch-four checkpoint is selected by validation VQA macro accuracy. Its archived released-model evaluation (0.00% VQA accuracy, 2.39% mIoU) predates the released-model (2.26% non-directional mIoU).

Building instance segmentation. OpenIns3D (Huang et al., 2024b) obtains AP_{25} and AP_{50} of 0.00% over the 16 test scenes because no building instances survive its fixed proposal and class-

Table 5: Representative scene-level VQA questions.

Building Reference	**Question Family**	**Example Question**
Absolute direction	Building distance	"How far is the southernmost building from the northwesternmost building? 0–5 meters, 5–10 meters, 10–15 meters, or 15 meters or more?"
Relative direction + feature	Attribute query	"What is the damage level of the southern single-family residential building in this scene?"
Two features	Comparison	"Is the building with a small amount of surrounding debris and a shingle roof more damaged than the building with a large amount of surrounding debris and a low-pitch roof?"

Table 6: Attribute accuracy (%). BT: building type; ST: stories; DL: damage level; FD: facade damage; RD: roof damage; RP: roof pitch; RT: roof type; SD: surrounding debris. The numbers of test labels, in this order, are 83, 85, 85, 85, 85, 84, 67, and 84.

Encoder	BT	ST	DL	FD	RD	RP	RT	SD
Paired images								
DINOv2-Base	98.80	100.00	63.53	76.47	67.06	73.81	76.12	36.90
SigLIP2-Base	100.00	100.00	61.18	78.82	71.76	48.81	83.58	42.86
OpenCLIP ViT-B/32	100.00	94.12	61.18	76.47	64.71	59.52	85.07	39.29
ConvNeXt-Base	93.98	97.65	62.35	75.29	44.71	55.95	83.58	30.95
ResNet-50	74.70	94.12	57.65	77.65	67.06	46.43	85.07	39.29
Majority	0.00	100.00	61.18	76.47	71.76	17.86	86.57	22.62
Point-cloud crops								
OpenShape	65.06	70.59	55.29	75.29	64.71	19.05	83.58	36.90
Sonata	32.53	80.00	10.59	32.94	51.76	17.86	86.57	22.62
Point-MAE	16.87	43.53	30.59	61.18	41.18	17.86	64.18	23.81
Majority	0.00	100.00	61.18	76.47	71.76	17.86	86.57	22.62

lookup thresholds. This reflects a failure of the evaluated configuration to produce accepted instances, not evidence that every possible proposal has zero overlap.

Table 7: Non-directional scene-task estimates and 95% scene-cluster bootstrap intervals (2,000 resamples, seed 13). Scores are exact-match accuracy for VQA and raw-point mIoU for segmentation; all values are percentages. Intervals are conditional on the available 16 scenes and do not represent training-seed variation.

Model	Score	95% interval
Point-cloud VQA		
Training-majority reference	44.88	39.07–49.08
3D-LLaVA	23.39	20.03–26.47
Grounded 3D-LLM	1.46	0.74–2.23
Rendered-scene VQA		
Training-majority reference	44.88	39.07–49.08
InternVL3-8B	23.68	18.36–28.40
LLaVA-OneVision-7B	29.53	25.52–33.37
Qwen2.5-VL-7B	37.28	33.39–40.85
Language-guided segmentation		
3D-LLaVA	1.41	0.85–2.08
Ground3D-LMM (released)	2.26	1.66–2.90
Grounded 3D-LLM	3.15	2.27–4.09
Reason3D	8.34	6.36–10.81

Table 8: Directional scene-task estimates and 95% scene-cluster bootstrap intervals (2,000 resamples, seed 13). Scores are exact-match accuracy for VQA and raw-point mIoU for segmentation; all values are percentages. Intervals are conditional on the available 16 scenes and do not represent training-seed variation.

Model	Score	95% interval
Point-cloud VQA		
3D-LLaVA	20.48	13.26–27.53
Training-majority reference	45.78	39.70–51.85
Rendered-scene VQA		
InternVL3-8B	13.86	8.45–19.27
LLaVA-OneVision-7B	14.16	9.37–18.53
Qwen2.5-VL-7B	13.55	8.46–18.53
Training-majority reference	45.78	39.70–51.85
Language-guided segmentation		
3D-LLaVA	1.58	0.86–2.34
Grounded 3D-LLM	1.81	1.14–2.80
Reason3D	5.30	4.33–6.08
Ground3D-LMM (released)	8.54	6.18–10.67