Holo-Captioning: Toward the Text Equivalent of 3D Scenes

Kun-Yu Lin[1], Chengke Bu[1], Zhenguo Li[2], and Kai Han[1⋆]

[1] Visual AI Lab, The University of Hong Kong
[2] Frontier Robotics

Abstract. This work introduces holo-captioning, a novel task that strives to seek the text equivalent of 3D scenes. As the initial step, we formulate holo-captioning as generating a structured textual description that comprehensively depicts all entities within a 3D scene—including their semantic tags, spatial locations, attributes, and inter-entity relations. To tackle this challenging task, we first develop an effective captioning engine to produce detailed descriptions of individual entity instances and instance pairs, and contribute a large-scale benchmark comprising over 15K scenes for training and evaluation. Building upon this foundation, we propose HoloScribe, a novel model that features an instance-aware decoupled pipeline for generating structured holo-captions, and further incorporates anchor-aware instance linking to identify relational instance pairs. Additionally, we propose a comprehensive evaluation metric named HoloScore, and provide a human-curated test set to ensure reliable model assessment. Experimental results demonstrate that HoloScribe significantly outperforms state-of-the-art 3D dense captioners and 3D LLM generalists, underscoring the effectiveness of our approach. Project page: https://visual-ai.github.io/holocap/

Keywords: 3D Scene Captioning · Holo-captioning

1 Introduction

Describing the 3D world with natural language is a fundamental field in computer vision and natural language processing, benefiting diverse 3D applications such as scene modeling [23, 71], vision navigation [65, 67] and robotic manipulation [5, 7, 83, 84]. As a pivotal task in this field, 3D dense captioning [13, 16] aims to simultaneously detect and describe object instances within a 3D scene. However, existing 3D dense captioning methods are constrained by a limited range of object categories and typically produce short, coarse textual descriptions. Driven by the rapid advances in Large Language Models (LLMs), some recent studies [3, 50] have proposed generating structured linguistic sequences to describe 3D scenes, thereby covering a broader range of categories in an open-vocabulary paradigm. Nevertheless, they primarily focus on detecting architectural layouts and salient objects, overlooking fine-grained entity attributes and inter-entity

⋆ Corresponding author.

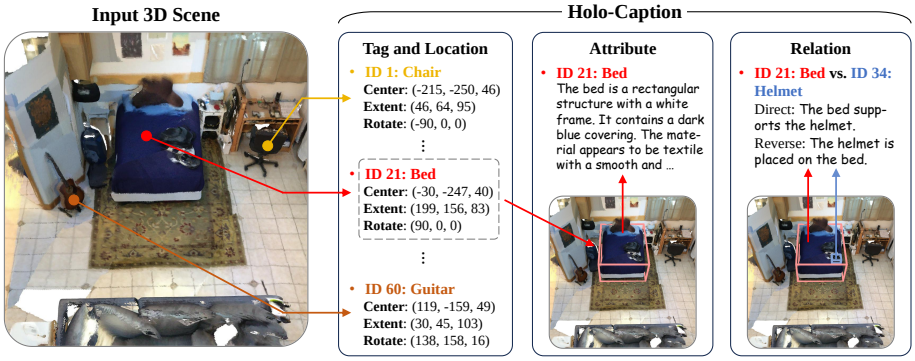

Fig. 1: An example to demonstrate our proposed *holo-captioning* task, which is formulated as generating a *comprehensive structured textual description* for a 3D scene. A holo-caption comprehensively depicts *all* entity instances within a 3D scene, including their respective *semantic tags*, *spatial locations* and *attributes*, as well as *relations* between entity instances. Best viewed in color.

relations. These omissions make existing methods inadequate for comprehensive scene understanding and limit their downstream applicability. Motivated by these limitations, this work conceives of finding the *text equivalent of a 3D scene*, an *ambitious yet very challenging* goal, which aims to develop a *comprehensive* textual description that captures all essential elements within a 3D scene, thereby offering great practical value and potentially facilitating diverse downstream applications.

In this work, we introduce holo-captioning, a novel task that strives to seek the text equivalent of 3D scenes. Our work serves as the *initial step* toward this challenging objective. To render the problem tractable, we formulate holo-captioning as generating a structured textual description of a 3D scene, which *comprehensively depicts all entity instances along with their semantic tags, spatial locations, attributes, and inter-entity relations*, as shown in Fig. 1. Compared with previous 3D dense captioners [13, 16] and structural description generators [3, 50], our formulation serves as a significantly closer approximation to the conceptual "text equivalence" due to its strong comprehensiveness. Specifically, it demands not only precise entity localization but also fine-grained descriptions for entity attributes and inter-entity relations. More importantly, our formulation requires models to directly predict entity locations *purely in textual form*, without relying on any auxiliary detectors or segmenters, thereby establishing a more intrinsic alignment between 3D scenes and text.

To systematically study holo-captioning, we propose HoloEngine, an effective captioning engine, and use it to construct HoloScan, a large-scale benchmark spanning over 15K indoor scenes across diverse categories. Specifically, HoloEngine produces high-quality captions in an instance-centric manner. First, given 3D oriented bounding boxes from a 3D scene, HoloEngine projects these boxes into each camera view via the known intrinsics and extrinsics, yielding view-

specific 2D boxes for each entity instance. Leveraging these view-specific boxes, we craft entity-aware prompts and instruct MLLMs to produce detailed view-specific descriptions for entity instances in perspective images. Finally, we use LLMs to perform viewpoint-aware consolidation, aggregating the view-specific texts into comprehensive captions. Built with HoloEngine, the HoloScan benchmark features comprehensive structured holo-captions and is accompanied by a test set rigorously curated by human experts to ensure reliability.

To support comprehensive evaluation of holo-captioning, we propose HoloScore, a tailored metric that assesses model performance across four dimensions. Traditional captioning metrics, typically based on n-gram or lexical similarity (*e.g.*, BLEU [52], CIDEr [57]), fall short in evaluating holo-captions due to their substantial length and high degree of granularity. To address this, HoloScore employs an LLM-empowered procedure, leveraging the structured nature of holo-captions. First, HoloScore performs grounded instance matching to assess semantic tagging and spatial localization. It then applies granular descriptor extraction to break long captions into atomic units. Finally, it evaluates the attribute and relation dimensions via dual descriptor matching.

Furthermore, we propose HoloScribe, a novel model to improve holo-captioning via an instance-aware decoupled pipeline. Our model is motivated by the fact that holo-captions are long and information-dense, making it difficult for models to perceive all elements within 3D scenes in a single forward pass. Accordingly, HoloScribe groups scene elements by entity instance and generates a structured textual description of the entire scene element by element. Specifically, HoloScribe first localizes entity instances and assigns semantic tags, and then generates detailed attribute and relation descriptions conditioned on the identified instances. To improve relation modeling, we further propose anchor-aware instance linking, which identifies relational instance pairs within a sampled subgraph conditioned on a given anchor instance. Our experiments demonstrate that HoloScribe significantly outperforms state-of-the-art 3D dense captioners and 3D LLM generalists in holo-captioning. Additionally, we evaluate HoloScribe in the 3D detection task on MultiScan [49], a dataset unseen during training. As shown in Table 1, it achieves superior performance over two representative 3D LLMs in a zero-shot setting. Furthermore, we deploy HoloScribe for downstream mobile robotic object navigation tasks, where it achieves a 75% success rate, surpassing the SpatialLM baseline (25%). These results underscore the effectiveness and practical utility of our task and model.

Table 1: Tagging and localization F-scores on MultiScan.

Models	Tag	Loc
LL3DA [15]	24.9	17.3
SpatialLM [50]	20.2	9.6
HoloScribe	53.5	20.2

In summary, our contributions are listed as follows: (1) We introduce the novel task of holo-captioning, which aims to find the textual equivalent of 3D scenes. We take the initial step toward this ambitious goal by formulating it as generating comprehensive structured textual descriptions that cover four fundamental dimensions, and we contribute a reliable metric for evaluation. (2) We propose an effective instance-centric captioning engine to produce high-quality holo-captions, and contribute a large-scale benchmark of

over 15K indoor scenes spanning diverse categories for training and evaluation. (3) We propose HoloScribe, a novel model that follows an instance-aware decoupled pipeline to generate comprehensive textual descriptions element by element. To the best of our knowledge, HoloScribe is the first 3D LLM capable of jointly localizing entity instances and generating detailed descriptions in *pure text form* without extra detectors.

2 Related Work

3D Captioning. 3D dense captioning [13] is a foundational task in the 3D vision-language field that aims to simultaneously detect and describe object instances within a 3D scene. As a pioneering work, Scan2Cap [13] proposed an end-to-end method that leverages attention mechanisms and message-passing graphs to process point clouds and generate captions. In recent years, the field has witnessed substantial progress driven by a variety of technical innovations [9, 14,16,17,36,38,77]. Despite these advances, existing models are built upon small-scale benchmarks (*e.g.*, ScanRefer [12], Nr3D [1]), which restrict the range of object categories and yield short, coarse descriptions. ExCap3D [76] represents a notable step forward, proposing a fine-grained captioning benchmark and a model capable of generating both detailed object- and part-level descriptions.

Overall, previous works typically use benchmarks with fewer than 1K scenes and rely on 3D detectors or segmenters for object localization [16,26,55]. *In contrast, our work covers a substantially larger number of scenes and directly predicts all elements (including entity locations) in pure text form without requiring any additional models, marking a significant stride toward 3D-text alignment.* Additionally, while holo-captioning is related to 3D object captioning, the latter focuses on generating descriptive sentences for isolated 3D objects [24,45,46,72] rather than complex multi-object scenes, and heavily relies on 2D MLLMs with limited 3D spatial capabilities. Another related field is 3D scene graph (3DSG) generation, which represents a 3D scene as a graph consisting of object nodes and relation edges [2,39,59,66]. 3DSG fundamentally differs from our task, as it primarily focuses on predicting objects and relations, whereas capturing other aspects such as attributes typically requires auxiliary models.

More broadly, our work pursues the *text equivalent of visual content*, a fundamental goal in vision-language research. Our prior panoptic captioning work [42] explores this goal for 2D images by representing an image as a purely textual description of entity instances, their locations and attributes, inter-entity relations, and the global image state. Holo-captioning extends this pursuit to 3D scenes, a more challenging setting where visual-text alignment shifts from the 2D image plane to metric 3D space, requiring entities to be characterized by their 3D positions, spatial extents, orientations, attributes, and physical relations. Together, panoptic captioning and holo-captioning are complementary steps toward our ultimate goal of visual-text equivalence.

3D Large Language Models (3D LLMs). Inspired by advances in 2D Multimodal LLMs (MLLMs) [22, 44, 60], the field of 3D LLMs [78] has witnessed re-

markable progress in recent years. These models develop general 3D scene understanding capabilities through training on diverse 3D vision-language tasks, such as 3D question answering and 3D grounding [1,4,12,30,31,34,43,47,48,61,74,80]. Existing 3D LLMs can be broadly categorized into two types based on their 3D vision representations. The first type relies on object-centric 3D representations [28, 29, 62, 85], *i.e.*, these models typically first parse objects using 3D detectors or segmenters, then extract 3D object features, and use these object features as vision inputs. The second type employs holistic scene-level representations without explicitly delineating objects [15,20,27,51,86], *e.g.*, LL3DA [15] develops a QFormer to connect 3D point clouds, visual prompts, and text. Additionally, motivated by the success of 2D MLLMs, some works attempt to construct 3D LLMs based on well-trained 2D video MLLMs, taking multi-view images as input [33, 53, 81]. The development of 3D LLMs can benefit a wide range of downstream tasks, such as robotic tasks [5, 7, 32, 35, 41, 63–65, 67–69]. Despite these advances, existing 3D LLMs are limited to handling relatively simple 3D vision-language tasks, *e.g.*, generating concise captions and recognizing a restricted set of object categories. Pioneering efforts have explored pure textual representations for 3D scenes by developing specialized 3D LLMs [3,50]. However, these works only focus on localizing and recognizing entity instances, while our holo-captioning aims to generate more comprehensive textual descriptions that incorporate attributes and relations. Another related work is Text-Scene [40], a hybrid approach that parses 3D scenes into textual descriptions using a suite of off-the-shelf models. This differs from our work, as we propose a single, unified model to produce comprehensive textual descriptions, with the goal of bridging 3D scenes and text.

3 Holo-Captioning

3.1 Task Formulation

In this work, we formulate holo-captioning as generating a structured textual description that *comprehensively* depicts all entity instances within a 3D scene, including their semantic tags, spatial locations, attributes, and the relations between entities. Unlike previous 3D captioning approaches [13, 16], our formulation encodes these four dimensions purely in text, resulting in a unified textual description that fully encapsulates the 3D scene. In principle, holo-captioning considers four dimensions as follows:

Semantic tag refers to the category label assigned to each entity instance in a 3D scene. We define "entities" as both architectural elements (*e.g.*, floor, wall, window) and free-standing objects (*e.g.*, table, bag, cabinet). Following SceneScript [3] and SpatialLM [50], we represent these category labels purely in text, without relying on class indices and predefined vocabularies.

Spatial location refers to the position and extent of an entity instance, represented by a 3D oriented bounding box. Specifically, in line with previous 3D object detection works [8,19,61], we represent the oriented bounding box of an entity instance as a 9-DoF vector $(c_x, c_y, c_z, e_x, e_y, e_z, r_x, r_y, r_z)$, where (c_x, c_y, c_z),

(e_x, e_y, e_z) and (r_x, r_y, r_z) denote the center, size, and Euler-angle orientation of the box, respectively. By using oriented bounding boxes, a holo-caption can accurately describe the locations and spatial extents of instances in pure text, avoiding vague expressions such as "at the center of the scene". As detailed in Sec. 5, our work adopts a structured text form to specify semantic tags and spatial locations of instances. Notably, the direct prediction of precise spatial coordinates in a purely textual form, without the aid of auxiliary models, is highly non-trivial, necessitating strong 3D-text alignment capabilities.

Attribute refers to characteristics or properties that describe the appearance or state of an entity instance. This dimension encompasses a wide range of attribute types, *e.g.*, color, shape, material, texture, and constituent parts. Our work uses free-form text to describe attributes, which facilitates fine-grained instance identification and enhances the comprehensiveness of scene descriptions [1,12,76].

Relation refers to the connections or interactions between two entity instances within a scene. This dimension encompasses a wide range of relation types, and our work focuses on those between nearby instances, as we primarily consider static indoor scenes where most meaningful relations occur among spatially adjacent entities. Typical examples include spatial relations (*e.g.*, "A is placed on B") and state relations (*e.g.*, "A is leaning against B"). Similar to the attribute dimension, we describe inter-entity relations in free-form text, which facilitates a more comprehensive understanding of the structural organization and contextual semantics of 3D scenes [11, 25, 54].

An example of our task formulation is presented in Fig. 1. By considering these four fundamental dimensions, holo-captioning leads to a comprehensive textual description of a 3D scene. Compared to previous works that pursue structural textual representations [3, 50], holo-captions not only capture entity categories and locations, but also describe entity attributes and inter-entity relations, thereby enabling a comprehensive understanding of 3D scenes. Notably, holo-captioning requires a single model to directly predict all scene elements *in pure text form* without relying on any additional models. This significantly differs from previous 3D dense captioning works that require auxiliary detectors to predict entity locations [13,16,17,76]. Such a formulation fosters a more intrinsic alignment between 3D scenes and text.

3.2 The HoloScore Metric

Previous 3D captioning works [13, 16] primarily rely on traditional metrics like BLEU [52], METEOR [6], and CIDEr [57]. However, these metrics are tailored for short captions and have been shown to perform poorly on long, descriptive texts [21,75]. To address this limitation, we propose HoloScore, a novel evaluation metric that comprehensively assesses holo-captions across four dimensions and achieves faithful alignment with our task objective.

Given the structured nature of holo-captions, we group caption elements into three aspects, namely entity tags with spatial locations, entity attributes, and inter-entity relations. Based on this grouping, we design a three-step evaluation

procedure comprising grounded instance matching, granular descriptor extraction, and dual descriptor matching. In what follows, we detail these three steps, and an overview of HoloScore is shown in Fig. 2.

Grounded Instance Matching. Given a set of ground-truth instances and a set of predicted ones, our grounded instance matching establishes a one-to-one correspondence between the two sets based on their semantic tags and spatial locations. Formally, let $\{(t_i, l_i)\}_{i=1}^{n}$ and $\{(\hat{t}_j, \hat{l}_j)\}_{j=1}^{m}$ denote the ground-truth and predicted entity instance sets, where n and m denote the total numbers of instances. Here, (t_i, l_i) represents the semantic tag and oriented bounding box of the i-th ground-truth instance, and $(\hat{t}_j, \hat{l}_j)$ represents those of the j-th prediction. Based on these definitions, we formulate grounded instance matching as an optimal bipartite matching problem:

$$\Theta = \arg\max_{\Theta} \sum_{i=1}^{n}\sum_{j=1}^{m}(\alpha\ \mathrm{sim}_{i,j} + \mathrm{iou}(l_i, \hat{l}_j)) \cdot \Theta_{i,j},\ \text{s.t.}\ \sum_{i} \Theta_{i,j} \leq 1,\ \sum_{j} \Theta_{i,j} \leq 1,$$

where Θ is a binary assignment matrix with $\Theta_{i,j} \in \{0, 1\}$ indicating whether the i-th ground-truth instance is matched to the j-th prediction, *i.e.*, $\Theta_{i,j} = 1$ means a match. $\mathrm{sim}_{i,j}$ denotes the semantic similarity between t_i and $\hat{t}_j$, computed using synonym matching and embedding similarity, while $\mathrm{iou}(l_i, \hat{l}_j)$ represents the 3D bounding-box IoU. A weighting factor $\alpha = 10$ is used to prioritize semantic tagging. In principle, two instances are considered a match when they share similar tags and are spatially close.

Based on the assignment matrix Θ, we quantify model performance in semantic tagging and spatial localization. For semantic tagging, a matched pair is regarded as semantically consistent if their category tags share synonymous nouns or have similar text embeddings, *i.e.*, $\mathrm{sim}_{i,j} \geq \delta_{\mathrm{tag}}$. By aggregating over all instances, we compute the precision and recall of semantic tagging and derive the corresponding F-score. Precision measures the proportion of correctly tagged predictions, while recall measures the proportion of ground-truth entities correctly identified. For spatial localization, two instances are regarded as location-consistent only if they are semantically consistent and their bounding boxes have an IoU above a preset threshold, *i.e.*, $\mathrm{iou}(l_i, \hat{l}_j) \geq \delta_{\mathrm{loc}}$. Following a similar protocol, we compute localization precision, recall, and the F-score to assess spatial localization quality.

Granular Descriptor Extraction. To evaluate the nuanced attribute dimension more precisely, we propose a granular descriptor extraction method to decompose long and detailed textual descriptions into a set of atomic descriptors. Specifically, for the i-th ground-truth entity instance, our method takes its paired attribute description A_i as input. We then prompt a state-of-the-art LLM to extract atomic descriptors from A_i, using carefully designed instruction prompts and in-context examples. This process is formulated as: $\{a_{i,k}\}_{k=1}^{n_i} = f_{\mathrm{gde}}(A_i)$, where $a_{i,k}$ denotes the k-th atomic descriptor, $f_{\mathrm{gde}}(\cdot)$ is the extraction function, and n_i is the number of extracted descriptors of the i-th instance. Similarly, the extracted descriptors of the j-th predicted instance are denoted by $\{\hat{a}_{j,k}\}_{k=1}^{m_j} = f_{\mathrm{gde}}(\hat{A}_j)$, where $\hat{A}_j$ denotes the predicted attribute description and

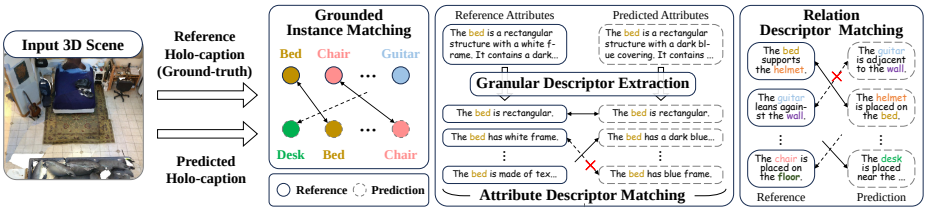

Fig. 2: An overview of our HoloScore metric. HoloScore first groups holo-caption elements into three aspects: semantic tags with spatial locations, entity attributes and inter-entity relations. It then performs grounded instance matching to assess tagging and localization, applies granular descriptor extraction to process attributes, and performs dual descriptor matching to assess attribute and relation dimensions.

m_j is the number of extracted descriptors. This decomposition yields superior interpretability and controllability compared to directly evaluating long, unstructured descriptions.

Dual Descriptor Matching. Finally, we evaluate model performance on the attribute and relation dimensions via descriptor matching. We employ an LLM-based embedding model [79] to encode text features, guided by a tailored prompt for robust similarity estimation between textual descriptors. For the attribute dimension, we establish a one-to-one correspondence between the atomic descriptors of each ground-truth instance and its matched predicted instance. The bipartite matching problem for the i-th ground-truth instance is formulated as follows:

$$\Omega^i = \arg\max_{\Omega^i} \sum_{k=1}^{n_i} \sum_{p=1}^{m_{\hat{i}}} \mathrm{sim}^i_{k,p} \cdot \Omega^i_{k,p}, \ \text{s.t.} \ \sum_k \Omega^i_{k,p} \leq 1, \ \sum_p \Omega^i_{k,p} \leq 1,$$

where Ω^i is the assignment matrix of the i-th instance, and $\Omega^i_{k,p} = 1$ indicates a match. $\mathrm{sim}^i_{k,p}$ denotes the feature similarity between descriptors $a_{i,k}$ and $\hat{a}_{\hat{i},p}$, where the $\hat{i}$-th predicted instance is the matched one. A descriptor pair is considered consistent when $\mathrm{sim}^i_{k,p} \geq \delta_{\mathrm{attr}}$. For each instance, we compute precision, recall, and the corresponding F-score, and then aggregate the F-scores across all instances to obtain the overall attribute metric.

For the relation dimension, we adopt a simplified procedure by direct descriptor matching, as relation descriptions are typically concise and each can be treated as a single atomic descriptor for efficiency. Given a ground-truth relation description $r_{i,j}$, we measure its similarity to the corresponding prediction $\hat{r}_{\hat{i},\hat{j}}$, where $\hat{i}$ and $\hat{j}$ are determined from the prior instance matching. A relation pair is considered consistent if its similarity exceeds δ_{rel}, and the overall F-score is derived in a similar manner to other dimensions.

In summary, our HoloScore metric comprises four F-scores, namely semantic tagging (s_t), spatial localization (s_l), entity attributes (s_a), and inter-entity relations (s_r). The final overall HoloScore is computed as $s = s_t + s_l + s_a + s_r$.

4 Captioning Engine and Benchmark

4.1 HoloEngine

According to the task definition, we aim to generate comprehensive holo-captions that are both highly structured and instance-isolated. To this end, we develop HoloEngine, an instance-centric captioning engine that produces structured descriptions, facilitating seamless integration across instances. Overall, HoloEngine first localizes entity instances in multi-view images based on projected 2D bounding boxes, then generates view-specific attribute and relation descriptions using state-of-the-art Multi-modal Large Language Models (MLLMs), and finally aggregates view-specific descriptions into holistic captions via LLMs.

Specifically, for the attribute dimension, we first obtain 3D oriented bounding boxes from 3D scenes, and project them into 2D regions across multi-view images using known camera intrinsics and extrinsics. Each 2D bounding box is overlaid on the image to explicitly indicate the target instance, and the category label is included in the text prompt (*e.g.*, "Describe the chair within the highlighted box"). This setup helps the MLLM focus on the correct instance and avoid confusion with nearby instances. The MLLM then produces detailed attribute descriptions, guided by prompts that discourage horizontal spatial terms (*e.g.*, "left", "right") and favor rotation-invariant phrasing (*e.g.*, "contain"), ensuring robustness to arbitrary scene orientations. Descriptions from all multi-view images are subsequently aggregated into a single, comprehensive caption by an LLM, conditioned on the corresponding viewing angles. To ensure efficiency, we sample entity-centric views at fixed angular intervals (*e.g.*, every 30°) instead of using all possible viewpoints.

For the relation dimension, we generate captions for instance pairs similarly, with three key differences. First, for each relational pair, we overlay a red and a blue bounding box in the image to explicitly indicate the subject and object instances, which we find significantly improves annotation quality in our preliminary validation. Second, as most inter-entity relations occur between nearby instances, HoloEngine focuses exclusively on spatially proximate instance pairs, thereby improving efficiency. Third, for each instance pair (*e.g.*, A and B), we instruct the MLLM to produce relation descriptions from both perspectives, namely A relative to B (*e.g.*, "A is on top of B") and B relative to A (*e.g.*, "B is below A"). This bidirectional strategy enriches the diversity of training data, and improves the comprehensiveness and reliability of evaluation. Moreover, to ensure higher annotation quality, we apply an MLLM-based category refinement step before MLLM captioning on both dimensions, assigning more specific category labels to instances with coarse ones (*e.g.*, replacing generic "object" labels).

4.2 HoloScan

Based on HoloEngine, we contribute a new benchmark named HoloScan for the holo-captioning task. We utilize publicly available 3D indoor scene scans from

five data sources, including four real datasets (ScanNet [18], 3RScan [58], Matterport3D [10] and ARKitScenes [19]) and one synthetic dataset Structured3D [82]. This results in a large-scale benchmark comprising over *8.4K real 3D scans* and 7.5K synthetic scenes. Its category labels and 3D bounding boxes are annotated with reference to EmbodiedScan [61], MMScan [47] and Structured3D. Following previous works [34,47,61], we standardize the data format, scene scale, sampling frequency, and viewpoint configuration across all datasets. In addition, we utilize only the spatial location information of Structured3D to improve training, as its synthetic instance appearance differs from that of real scenes. During data production, we select the best MLLM for each step through preliminary experiments, balancing image understanding, instruction-following, and inference efficiency. Overall, HoloScan comprises over 13K training scenes and 619 validation scenes paired with high-quality auto-generated holo-captions, along with 83 test scenes paired with holo-captions rigorously curated by human experts.

Table 2: Comparison with 3D dense captioning and structural description generation benchmarks.

Data	Type	Box	Category	Scene	Word
ScanRefer [12]	Real	✓	18	0.7K	17.9
Nr3D [1]	Real	✓	18	0.6K	11.4
ExCap3D [76]	Real	✗	84	0.9K	82.8
ASE [3]	Synthetic	✓	Layout	100K	✗
SpatialLM [50]	Synthetic	✓	62	54K	✗
HoloScan	Hybrid	✓	**734**	15K	**89.0**

In summary, HoloScan features high-quality structured holo-captions that comprehensively depict scene elements along four fundamental dimensions. As shown in Table 2, compared with previous 3D dense captioning benchmarks [1,12,76], HoloScan provides longer and more detailed instance descriptions (more words on average) and encompasses a broader range of entity categories. Moreover, it covers a wider variety of *real scenes* and sources data from multiple datasets, unlike previous benchmarks that rely solely on a single source (*e.g.*, ScanNet [18]). Additionally, unlike previous structural description generation benchmarks that focus primarily on *synthetic* scenes [3,50], HoloScan contains a substantial number of *real-world scenes* and provides *comprehensive descriptions* with fine-grained coverage of attributes and relations. Collectively, these advancements position our task formulation as a significantly closer approximation to the conceptual "text equivalence".

5 HoloScribe

To address holo-captioning, we propose HoloScribe, a novel model that follows an instance-aware decoupled pipeline. The design of HoloScribe is motivated by the fact that holo-captions are long and information-dense, making it difficult for models to perceive all elements in 3D scenes in a single forward pass. To tackle this challenge, HoloScribe decomposes the captioning process into instance-centric subtasks, progressively generating the structured textual description of a scene element by element. Specifically, HoloScribe first localizes entity instances and assigns semantic tags, and then generates detailed attribute

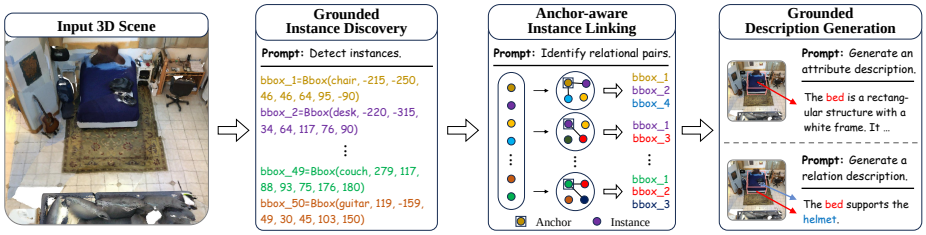

Fig. 3: An overview of our HoloScribe. Given an input 3D scene, HoloScribe generates holo-captions following an instance-aware decoupling pipeline, consisting of grounded instance discovery, anchor-aware instance linking, and grounded description generation.

and relation descriptions conditioned on the identified instances. An overview of HoloScribe is shown in Fig. 3.

Formally, HoloScribe takes as input a 3D scene in point cloud format, denoted by $\mathbf{Q}_v$, and produces a holo-caption $\hat{\mathbf{H}}$ via three phases, namely grounded instance discovery, anchor-aware instance linking, and grounded description generation, which we detail below. Following SpatialLM [50], HoloScribe adopts a LLaVA-style 3D LLM architecture, consisting of a point cloud encoder, an MLP connector, and a decoder-only LLM. In each phase, we employ a specific prompt to guide HoloScribe in producing desired outputs in pure text form.

Grounded Instance Discovery. This phase aims to identify all entity instances within the input scene, serving as the foundation for a holo-caption. For training, we first extract entity instances along with their semantic tags and spatial locations from the ground-truth holo-caption $\mathbf{H}$. We then construct a point-text training tuple $(\mathbf{Q}_v, \mathbf{Q}_{\text{inst}}, \mathbf{H}_{\text{inst}})$, where $\mathbf{Q}_{\text{inst}}$ denotes the tailored prompt and $\mathbf{H}_{\text{inst}}$ is a holo-instance text composed of individual instance texts. During inference, this phase outputs a predicted holo-instance text $\hat{\mathbf{H}}_{\text{inst}}$ given a scene. Each instance in $\mathbf{H}_{\text{inst}}$ is represented by a structured text of the format "`bbox_`i`=Bbox(`t_i`,`$c_{i,x}$`,`$c_{i,y}$`,`$c_{i,z}$`,`$e_{i,x}$`,`$e_{i,y}$`,`$e_{i,z}$`,`$r_{i,z}$`)`", where i denotes the instance index, t_i is the semantic tag, and the remaining terms are box parameters as defined in Sec. 3.1. To simplify learning, we adopt 7-DoF bounding boxes during training, which are recomputed from each entity's point cloud by constraining $r_{i,x} = r_{i,y} = 0$. Given that real-world scenes often contain dozens of instances, we further divide them into three types, namely *architectural elements*, *common objects*, and *other objects*. During inference, we predict each type independently with tailored prompts and then aggregate the results.

Anchor-aware Instance Linking. Given the detected entity instances, this phase aims to identify instance pairs that exhibit inter-entity relations. A straightforward approach is to predict relations for every instance pair in the scene. However, since typical scenes often contain dozens of instances, exhaustively considering every possible pair would result in an excessive number of relation candidates and high computational cost. To address this, we propose an anchor-aware edge formation strategy that learns relation modeling within sampled subgraphs of the complete instance graph. This design enables HoloScribe to

efficiently focus on instance pairs that are likely to exhibit meaningful relationships, and to generate relation descriptions only for these selected pairs.

Given a scene, we first construct a graph where each entity instance corresponds to a node. The goal of anchor-aware instance linking is to determine whether an edge should be formed between two nodes, indicating the presence of a relationship between the corresponding instances. Since the complete instance graph can be large, we instead sample smaller subgraphs and perform instance linking within these subgraphs to make relation learning more tractable. During training, we sample a large and diverse set of subgraphs to improve learning coverage. This yields a set of point-text tuples $\{(\mathbf{Q}_v, \mathbf{Q}_{\text{pair}}, \mathbf{H}_{\text{subg}}, \mathbf{H}_{\text{pair}})\}$ from the ground-truth instance set, where $\mathbf{Q}_{\text{pair}}$ is the prompt, $\mathbf{H}_{\text{subg}}$ is text describing a sampled subgraph, and $\mathbf{H}_{\text{pair}}$ is a relation text specifying relationship existence.

For each subgraph, we designate one instance as the anchor and sample $n_g - 1$ additional instances to form a subgraph with n_g nodes in total. The subgraph is then serialized into $\mathbf{H}_{\text{subg}}$ and fed into HoloScribe. Its textual format mirrors the holo-instance text used in grounded instance discovery, listing the n_g sampled instances with the anchor instance always placed first. The corresponding target text $\mathbf{H}_{\text{pair}}$ specifies which of the remaining instances are related to the anchor. For example, the relation text "`bbox_1,bbox_2,bbox_3`" indicates that the second and third instances are related to the anchor (*i.e.*, the first instance). To ensure decoding stability and prevent empty outputs, we enforce that each relation text always begins with "`bbox_1`". During inference, each discovered instance is treated as the anchor in turn, and we construct subgraphs $\{\hat{\mathbf{H}}_{\text{subg}}\}$ accordingly. By aggregating the predicted relation texts $\{\hat{\mathbf{H}}_{\text{pair}}\}$ across all anchor-centric subgraphs, we obtain the complete set of relational instance pairs for the entire scene.

Grounded Description Generation. After discovering entity instances and identifying relational pairs, we proceed to generate detailed attribute and relation descriptions based on entity tags and locations. For the attribute dimension, we construct point-text tuples $\{(\mathbf{Q}_v, \mathbf{Q}_{\text{attr}}, \mathbf{H}_i, \mathbf{H}_i^a)\}$ for training, where $\mathbf{Q}_{\text{attr}}$ is the prompt. $\mathbf{H}_i$ is a text indicating the semantic tag and spatial location of the i-th ground-truth instance, resembling holo-instance texts. $\mathbf{H}_i^a$ is the corresponding ground-truth attribute description, which serves as the target output of the attribute dimension. Similarly, for the relation dimension, we construct point-text tuples $\{(\mathbf{Q}_v, \mathbf{Q}_{\text{rel}}, \mathbf{H}_{i,j}, \mathbf{H}_{i,j}^r)\}$ for training. $\mathbf{Q}_{\text{rel}}$ is the prompt, $\mathbf{H}_{i,j}$ is the instance text composed of the i-th and j-th instances, and $\mathbf{H}_{i,j}^r$ is the corresponding ground-truth relation description. During inference, we take the discovered instances $\{\hat{\mathbf{H}}_i\}$ and identified instance pairs $\{\hat{\mathbf{H}}_{i,j}\}$ as inputs, and generate attribute descriptions $\{\hat{\mathbf{H}}_i^a\}$ and relation descriptions $\{\hat{\mathbf{H}}_{i,j}^r\}$. Finally, by aggregating model outputs across all four dimensions, HoloScribe produces complete holo-captions in a structured text form.

Table 3: Comparison results on the validation and test sets of HoloScan. Performance is measured by HoloScore, and we report the scores in tagging, location, attribute, relation, and the overall score. **Bold** indicates the best, and "-" means the model does not support that dimension.

Models	LLM	Tagging	Location	Attribute	Relation	Overall
Validation Set						
Vote2Cap-DETR [16]	-	32.87	17.00	2.49	-	-
Vote2Cap-DETR++ [17]	-	30.84	16.70	2.38	-	-
LEO [29]	Vicuna-7B	23.34	14.25	1.53	-	-
LL3DA [15]	OPT-1.3B	31.28	16.50	2.99	-	-
SpatialLM-Tuned [29]	Qwen2.5-0.5B	51.73	9.94	6.74	2.60	71.01
LL3DA-Tuned [15]	Qwen2.5-0.5B	31.25	16.50	10.36	4.83	62.94
HoloScribe (Ours)	Qwen2.5-0.5B	**60.43**	**22.33**	**22.52**	**11.08**	**116.36**
Test Set						
Vote2Cap-DETR [16]	-	32.50	17.71	2.70	-	-
Vote2Cap-DETR++ [17]	-	31.39	17.04	2.67	-	-
LEO [29]	Vicuna-7B	25.91	14.93	1.60	-	-
LL3DA [15]	OPT-1.3B	31.28	16.89	3.27	-	-
SpatialLM-Tuned [29]	Qwen2.5-0.5B	49.82	9.23	7.01	3.57	69.63
LL3DA-Tuned [15]	Qwen2.5-0.5B	31.28	16.88	9.54	3.97	61.67
HoloScribe (Ours)	Qwen2.5-0.5B	**61.83**	**22.75**	**23.72**	**8.74**	**117.04**

6 Experiment

Implementation Details. Following SpatialLM [50], our HoloScribe adopts an "Encoder-MLP-LLM" architecture with Sonata [70] as the vision encoder, and it is initialized using the SpatialLM1.1-Qwen-0.5B checkpoint. We finetune our model on the training set of HoloScan for one epoch using LoRA. For HoloEngine, we use Gemma3-12B [37] to produce descriptions of images and Qwen3-4B-Thinking [73] to aggregate descriptions. For HoloScore, we use Gemma3-1B [37] for granular descriptor extraction, and use Qwen3-Embedding-0.6B [79] for encoding text in dual descriptor matching. The coefficients are set as $\delta_{\text{tag}} = 0.5$, $\delta_{\text{loc}} = 0.3$, and $\delta_{\text{attr}} = \delta_{\text{rel}} = 0.6$.

Comparison with State-of-the-Arts. Table 3 summarizes the comparison with state-of-the-art captioning models, including 3D dense captioners [16, 17], 3D LLM generalists [15, 29], and holo-captioning specialists (see the Appendix for baseline details). From the table, we find that both 3D dense captioners and 3D LLM generalists obtain poor performance in holo-captioning, especially on the attribute dimension. This is because these models are trained on traditional captioning benchmarks (*e.g.*, ScanRefer [12]) and thus can only produce short captions describing simple attributes like color or shape. Compared with holo-captioning specialists trained on HoloScan, our HoloScribe obtains significant improvements on all dimensions, which is attributed to our proposed instance-aware decoupled pipeline.

Ablation Study. Table 4 summarizes the ablation study of HoloScribe. We adopt the directly tuned SpatialLM as the baseline, which performs poorly be-

Table 4: Ablation study of our HoloScribe on the validation set of HoloScan.

Models	Tagging	Location	Attribute	Relation	Overall
Baseline	46.98	8.23	2.39	1.04	58.64
w/ decoupling	51.73	9.94	6.74	2.60	71.01
w/ instance-aware prediction	59.20	20.45	22.41	7.92	109.98
Full w/o anchor	58.67	17.94	22.46	9.20	108.27
Full	**60.43**	**22.33**	**22.52**	**11.08**	**116.36**

cause the desired holo-captioning outputs are long and highly content-rich. After isolating the relation prediction component, we observe a clear performance improvement, as shown by "w/ decoupling". By introducing the instance-aware attribute and relation prediction mechanism ("w/ instance-aware prediction"), the overall performance of HoloScribe improves significantly, where relation prediction is performed for every instance pair. By further incorporating our anchor-aware linking mechanism ("Full"), HoloScribe obtains substantial improvement on the relation dimension, as the model can identify meaningful relational pairs while filtering out irrelevant pairs, thereby reducing erroneous predictions. Finally, we conduct an additional comparison against a variant without the anchor mechanism ("Full w/o anchor"), *i.e.*, one that predicts all relational pairs within each subgraph. The results demonstrate the effectiveness of our anchor-aware design, as the anchor-free variant struggles to capture relations among all possible pairs simultaneously.

3D Scene Reconstruction. We conduct a scene reconstruction experiment to qualitatively demonstrate the effectiveness and value of our model. Specifically, we employ Hunyuan3D [56] to generate 3D objects using instance tags and attributes derived from a generated holo-caption. The size and pose of each object are then refined according to the predicted bounding boxes, and the objects are positioned in 3D space to assemble the reconstructed scene. As shown in Fig. 4 (a) and (b), HoloScribe produces reconstructed 3D scenes that closely match the original ones, particularly in terms of accurate semantic tagging and precise entity localization. As shown in Fig. 4 (c), our reconstruction outperforms the LL3DA-based one, as LL3DA suffers from inaccurate entity detection (yielding redundant, incorrect chairs) and monotonous attributes (usually predicting the single "black" attribute).

Cross-Dataset Detection. To demonstrate the generalizability of HoloScribe, we conduct a detection experiment using MultiScan [49], which is unseen during training. Specifically, for each scene, we apply HoloScribe to produce a holo-caption, from which we extract the entity tags and locations to perform the detection task. As shown in Table 1, two representative 3D LLMs, namely LL3DA and SpatialLM, perform poorly on MultiScan. This is because LL3DA can only detect a limited range of categories, and SpatialLM is trained solely on synthetic scenes and thus hardly generalizes to real scenes. In contrast, HoloScribe significantly outperforms these baselines as it can recognize diverse categories

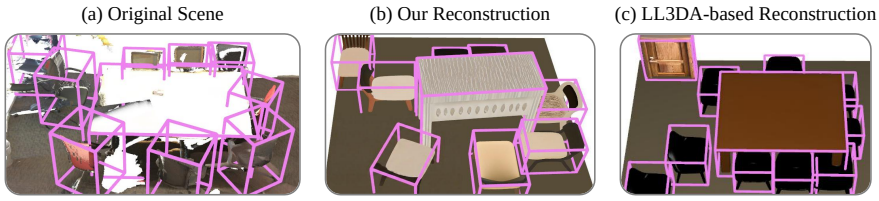

Fig. 4: Demonstration of (a) a real scene from HoloScan, (b) the HoloScribe-based reconstruction, and (c) LL3DA-based reconstruction. We highlight entities by 3D boxes.

and generalize more effectively to real scenes from unseen datasets. Additional results in the Appendix show that holo-captions enable flexible scene editing.

7 Conclusion

This work introduces holo-captioning, a novel task that pursues the text equivalent of 3D scenes. We take the initial step toward this ambitious goal by formulating it as the generation of a structured textual description that comprehensively depicts scene elements across four dimensions. To enable systematic study, we propose a comprehensive evaluation metric, develop an effective captioning engine, and contribute a large-scale benchmark. We further propose HoloScribe, which follows an instance-aware decoupled pipeline to address holo-captioning. To our knowledge, HoloScribe is the first 3D LLM capable of jointly localizing entity instances and generating detailed descriptions purely in text, demonstrating strong performance in extensive experiments.

Despite our efforts, this work represents only the initial attempt toward our conceptual goal of "text equivalence", and holo-captioning still presents numerous open challenges. Our experiments reveal that the task remains highly challenging for existing models, partially due to the inherent difficulty of predicting entity locations in pure text form. We consider scaling up the model a promising direction for future work. Nevertheless, this work establishes a solid foundation for future research and is expected to catalyze further advances in bridging 3D scenes and text.

Acknowledgments. This work is supported by the Hong Kong Research Grants Council - General Research Fund (Grant No.: 17211024), the Innovation and Technology Fund (ITS/488/24FP), and the HKU Seed Fund for PI Research. The authors would like to thank Yi-Xiang He and Yi-Lin Wei for their support in downstream experiments. The authors also thank Jiaming Zhou, Yu-Ming Tang, and Weining Ren for their valuable suggestions on the writing. The authors also thank Yuxian Li, Mingshan Huang, Huairong Chen, Shenru Zhang, Yixuan Chen, Junming Chen, Yifei Zhang, Zhenyi Fan, Jialu Tang, and Boning Shao for their help and support for our project.

References

1. Achlioptas, P., Abdelreheem, A., Xia, F., Elhoseiny, M., Guibas, L.J.: ReferIt3D: Neural listeners for fine-grained 3D object identification in real-world scenes. In: European Conference on Computer Vision (2020) 4, 5, 6, 10
2. Armeni, I., He, Z., Zamir, A., Gwak, J., Malik, J., Fischer, M., Savarese, S.: 3D Scene Graph: A structure for unified semantics, 3D space, and camera. In: IEEE/CVF International Conference on Computer Vision (2019) 4
3. Avetisyan, A., Xie, C., Howard-Jenkins, H., Yang, T., Aroudj, S., Patra, S., Zhang, F., Frost, D.P., Holland, L., Orme, C., Engel, J.J., Miller, E., Newcombe, R.A., Balntas, V.: SceneScript: Reconstructing scenes with an autoregressive structured language model. In: European Conference on Computer Vision (2024) 1, 2, 5, 6, 10
4. Azuma, D., Miyanishi, T., Kurita, S., Kawanabe, M.: ScanQA: 3D question answering for spatial scene understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022) 5
5. Bai, S., Song, W., Chen, J., Ji, Y., Zhong, Z., Yang, J., Zhao, H., Zhou, W., Zhao, W., Li, Z., Ding, P., Chi, C., Li, H., Xu, C., Zheng, X., Wang, D., Zhang, S., Chen, B.: Towards a unified understanding of robot manipulation: A comprehensive survey. arXiv preprint arXiv:2510.10903 (2025) 1, 5
6. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005 (2005) 6
7. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., Jakubczak, S., Jones, T., Ke, L., Levine, S., Li-Bell, A., Mothukuri, M., Nair, S., Pertsch, K., Shi, L.X., Tanner, J., Vuong, Q., Walling, A., Wang, H., Zhilinsky, U.: π_0: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164 (2024) 1, 5
8. Brazil, G., Kumar, A., Straub, J., Ravi, N., Johnson, J., Gkioxari, G.: Omni3D: A large benchmark and model for 3D object detection in the wild. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023) 5
9. Cai, D., Zhao, L., Zhang, J., Sheng, L., Xu, D.: 3DJCG: A unified framework for joint dense captioning and visual grounding on 3D point clouds. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16464–16473 (2022) 4
10. Chang, A., Dai, A., Funkhouser, T., Halber, M., Nießner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3D: Learning from rgb-d data in indoor environments. In: International Conference on 3D Vision (2017) 10
11. Chang, H., Boyalakuntla, K., Lu, S., Cai, S., Jing, E.P., Keskar, S., Geng, S., Abbas, A., Zhou, L., Bekris, K.E., Boularias, A.: Context-aware entity grounding with open-vocabulary 3D scene graphs. In: Conference on Robot Learning (2023) 6
12. Chen, D.Z., Chang, A.X., Nießner, M.: ScanRefer: 3D object localization in RGB-D scans using natural language. In: European Conference on Computer Vision (2020) 4, 5, 6, 10, 13
13. Chen, D.Z., Gholami, A., Nießner, M., Chang, A.X.: Scan2Cap: Context-aware dense captioning in RGB-D scans. In: IEEE Conference on Computer Vision and Pattern Recognition (2021) 1, 2, 4, 5, 6

14. Chen, D.Z., Wu, Q., Nießner, M., Chang, A.X.: D3Net: A unified speaker-listener architecture for 3D dense captioning and visual grounding. In: European Conference on Computer Vision. Springer-Verlag, Berlin, Heidelberg (2022) 4
15. Chen, S., Chen, X., Zhang, C., Li, M., Yu, G., Fei, H., Zhu, H., Fan, J., Chen, T.: LL3DA: visual interactive instruction tuning for omni-3D understanding, reasoning, and planning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024) 3, 5, 13
16. Chen, S., Zhu, H., Chen, X., Lei, Y., Yu, G., Chen, T.: End-to-end 3D dense captioning with vote2cap-detr. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023) 1, 2, 4, 5, 6, 13
17. Chen, S., Zhu, H., Li, M., Chen, X., Guo, P., Lei, Y., Yu, G., Li, T., Chen, T.: Vote2Cap-DETR++: Decoupling localization and describing for end-to-end 3D dense captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence **46** (2024) 4, 6, 13
18. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (2017) 10
19. Dehghan, A., Baruch, G., Chen, Z., Feigin, Y., Fu, P., Gebauer, T., Kurz, D., Dimry, T., Joffe, B., Schwartz, A., Shulman, E.: ARKitScenes: A diverse real-world dataset for 3D indoor scene understanding using mobile RGB-D data. In: Neural Information Processing Systems Datasets and Benchmarks Track (2021) 5, 10
20. Deng, J., He, T., Jiang, L., Wang, T., Dayoub, F., Reid, I.D.: 3D-LLaVA: Towards generalist 3D lmms with omni superpoint transformer. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025) 5
21. Dong, H., Li, J., Wu, B., Wang, J., Zhang, Y., Guo, H.: Benchmarking and improving detail image caption. arXiv preprint arXiv:2405.19092 (2024) 6
22. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024) 4
23. Fang, C., Li, H., Liang, Y., Zheng, J., Mao, Y., Liu, Y., Tang, R., Zhou, Z., Tan, P.: SPATIALGEN: Layout-guided 3D indoor scene generation. arXiv preprint arXiv:2509.14981 (2025) 1
24. Ge, Y., Zeng, X., Huffman, J.S., Lin, T., Liu, M., Cui, Y.: Visual Fact Checker: Enabling high-fidelity detailed caption generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024) 4
25. Gu, Q., Kuwajerwala, A., Morin, S., Jatavallabhula, K.M., Sen, B., Agarwal, A., Rivera, C., Paul, W., Ellis, K., Chellappa, R., Gan, C., de Melo, C.M., Tenenbaum, J.B., Torralba, A., Shkurti, F., Paull, L.: ConceptGraphs: Open-vocabulary 3D scene graphs for perception and planning. In: IEEE International Conference on Robotics and Automation (2024) 6
26. Guo, K., Huang, Y., Jia, T., Song, X., Sun, S., Wei, H., Han, X.F., Huang, S., Strisciuglio, N., Li, S.: Visual Grounding in 2D and 3D: A unified perspective and survey. Inf. Fusion (2026) 4
27. Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., Gan, C.: 3D-LLM: Injecting the 3D world into large language models. In: Neural Information Processing Systems (2023) 5
28. Huang, H., Chen, Y., Wang, Z., Huang, R., Xu, R., Wang, T., Liu, L., Cheng, X., Zhao, Y., Pang, J., Zhao, Z.: Chat-Scene: Bridging 3D scene and large language models with object identifiers. In: Neural Information Processing Systems (2024) 5

29. Huang, J., Yong, S., Ma, X., Linghu, X., Li, P., Wang, Y., Li, Q., Zhu, S., Jia, B., Huang, S.: An embodied generalist agent in 3D world. In: International Conference on Machine Learning (2024) 5, 13
30. Huang, J., Li, Z., Zhang, H., Chen, R., He, X., Guo, Y., Wang, W., Liu, T., Gong, M.: SURPRISE3D: A dataset for spatial understanding and reasoning in complex 3D scenes. arXiv preprint arXiv:2507.07781 (2025) 5
31. Huang, T., Zhang, Z., Tang, H.: 3D-R1: Enhancing reasoning in 3D VLMs for unified scene understanding. arXiv preprint arXiv:2507.23478 (2025) 5
32. Huang, W.J., Zhang, Y.Y., Wei, Y.L., Xia, Z.W., Tan, J., Li, Y.M., Zhao, Z., Zheng, W.S.: Beyond Mimicry: Learning whole-body human-humanoid interaction from human-human demonstrations. In: IEEE Conference on Computer Vision and Pattern Recognition (2026) 5
33. Huang, X., Wu, J., Xie, Q., Han, K.: 3DRS: MLLMs Need 3D-Aware Representation Supervision for Scene Understanding. In: Neural Information Processing Systems (2025) 5
34. Jia, B., Chen, Y., Yu, H., Wang, Y., Niu, X., Liu, T., Li, Q., Huang, S.: SceneVerse: Scaling 3D vision-language learning for grounded scene understanding. In: European Conference on Computer Vision (2024) 5, 10
35. Jiang, J., Wu, X., He, Y., Zeng, L., Wei, Y., Zhang, D., Zheng, W.: Rethinking bimanual robotic manipulation: Learning with decoupled interaction framework. In: IEEE/CVF International Conference on Computer Vision (2025) 5
36. Jiao, Y., Chen, S., Jie, Z., Chen, J., Ma, L., Jiang, Y.G.: MORE: Multi-order relation mining for dense captioning in 3D scenes. In: European Conference on Computer Vision. pp. 528–545. Springer Nature Switzerland, Cham (2022) 4
37. Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., et al.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025) 13
38. Kim, M., Lim, H., Kim, S.H., Lee, S., Kim, B., Kim, G.: See It All: Contextualized late aggregation for 3D dense captioning. In: Findings of the Association for Computational Linguistics (2024) 4
39. Koch, S., Vaskevicius, N., Colosi, M., Hermosilla, P., Ropinski, T.: Open3DSG: Open-vocabulary 3D scene graphs from point clouds with queryable objects and open-set relationships. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024) 4
40. Li, H., Liu, R., Fan, H., Yang, Y.: Text-scene: A scene-to-language parsing framework for 3D scene understanding. arXiv preprint arXiv:2509.16721 (2025) 5
41. Li, Y.M., Huang, W.J., Wang, A.L., Zeng, L.A., Meng, J.K., Zheng, W.S.: EgoExo-Fitness: Towards egocentric and exocentric full-body action understanding. In: European Conference on Computer Vision. Springer (2024) 5
42. Lin, K.Y., Wang, H., Ren, W., Han, K.: Panoptic Captioning: An equivalence bridge for image and text. In: Neural Information Processing Systems (2025) 4
43. Linghu, X., Huang, J., Niu, X., Ma, X., Jia, B., Huang, S.: Multi-modal situated reasoning in 3D scenes. In: Neural Information Processing Systems Datasets and Benchmarks Track (2024) 5
44. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Neural Information Processing Systems (2023) 4
45. Luo, T., Johnson, J., Lee, H.: View selection for 3D captioning via diffusion ranking. In: European Conference on Computer Vision (2024) 4
46. Luo, T., Rockwell, C., Lee, H., Johnson, J.: Scalable 3D captioning with pretrained models. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Neural Information Processing Systems (2023) 4

47. Lyu, R., Lin, J., Wang, T., Yang, S., Mao, X., Chen, Y., Xu, R., Huang, H., Zhu, C., Lin, D., Pang, J.: MMScan: A multi-modal 3D scene dataset with hierarchical grounded language annotations. In: Neural Information Processing Systems (2024) 5, 10
48. Ma, X., Yong, S., Zheng, Z., Li, Q., Liang, Y., Zhu, S.C., Huang, S.: SQA3D: Situated question answering in 3D scenes. In: International Conference on Learning Representations (2023) 5
49. Mao, Y., Zhang, Y., Jiang, H., Chang, A.X., Savva, M.: MultiScan: Scalable RGBD scanning for 3D environments with articulated objects. In: Neural Information Processing Systems (2022) 3, 14
50. Mao, Y., Zhong, J., Fang, C., Zheng, J., Tang, R., Zhu, H., Tan, P., Zhou, Z.: SpatialLM: Training large language models for structured indoor modeling. In: Neural Information Processing Systems (2025) 1, 2, 3, 5, 6, 10, 11, 13
51. Mei, G., Lin, W., Riz, L., Wu, Y., Poiesi, F., Wang, Y.: PerLA: Perceptive 3D language assistant. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025) 5
52. Papineni, K., Roukos, S., Ward, T., Zhu, W.: Bleu: a method for automatic evaluation of machine translation. In: Annual Meeting of the Association for Computational Linguistics (2002) 3, 6
53. Qi, Z., Zhang, Z., Fang, Y., Wang, J., Zhao, H.: GPT4Scene: Understand 3D scenes from videos with vision-language models. arXiv preprint arXiv:2501.01428 (2025) 5
54. Rana, K., Haviland, J., Garg, S., Abou-Chakra, J., Reid, I.D., Sünderhauf, N.: SayPlan: Grounding large language models using 3D scene graphs for scalable robot task planning. In: Conference on Robot Learning (2023) 6
55. Schult, J., Engelmann, F., Hermans, A., Litany, O., Tang, S., Leibe, B.: Mask3D: Mask transformer for 3D semantic instance segmentation. In: IEEE International Conference on Robotics and Automation (2023) 4
56. Team Hunyuan3D, Yang, S., Yang, M., Feng, Y., Huang, X., et al.: Hunyuan3D 2.1: From images to high-fidelity 3D assets with production-ready PBR material. arXiv preprint arXiv:2506.15442 (2025) 14
57. Vedantam, R., Zitnick, C.L., Parikh, D.: CIDEr: Consensus-based image description evaluation. In: IEEE Conference on Computer Vision and Pattern Recognition (2015) 3, 6
58. Wald, J., Avetisyan, A., Navab, N., Tombari, F., Niessner, M.: RIO: 3D object instance re-localization in changing indoor environments. In: IEEE/CVF International Conference on Computer Vision (2019) 10
59. Wald, J., Dhamo, H., Navab, N., Tombari, F.: Learning 3D semantic scene graphs from 3D indoor reconstructions. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020) 4
60. Wang, A., Shan, B., Shi, W., Lin, K., Fei, X., Tang, G., Liao, L., Tang, J., Huang, C., Zheng, W.: ParGo: Bridging vision-language with partial and global views. In: AAAI (2025) 4
61. Wang, T., Mao, X., Zhu, C., Xu, R., Lyu, R., Li, P., Chen, X., Zhang, W., Chen, K., Xue, T., Liu, X., Lu, C., Lin, D., Pang, J.: EmbodiedScan: A holistic multi-modal 3D perception suite towards embodied AI. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024) 5, 10
62. Wang, Z., Huang, H., Zhao, Y., Zhang, Z., Zhao, Z.: Chat-3D: Data-efficiently tuning large language model for universal dialogue of 3D scenes. arXiv preprint arXiv:2308.08769 (2023) 5

63. Wei, Y.L., Jiang, J.J., Xing, C., Tan, X.T., Wu, X.M., Li, H., Cutkosky, M., Zheng, W.S.: Grasp as you say: Language-guided dexterous grasp generation. In: Neural Information Processing Systems (2024) 5
64. Wei, Y.L., Lin, M., Lin, Y., Jiang, J.J., Wu, X.M., Zeng, L.A., Zheng, W.S.: AffordDexGrasp: Open-set language-guided dexterous grasp with generalizable-instructive affordance. In: IEEE/CVF International Conference on Computer Vision. pp. 11818–11828 (2025) 5
65. Werby, A., Huang, C., Büchner, M., Valada, A., Burgard, W.: Hierarchical open-vocabulary 3D scene graphs for language-grounded robot navigation. In: Robotics: Science and Systems (2024) 1, 5
66. Wu, S., Fei, H., Chua, T.: Universal scene graph generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025) 4
67. Wu, W., Chang, T., Li, X., Yin, Q., Hu, Y.: Vision-language navigation: a survey and taxonomy. Neural Comput. Appl. (2024) 1, 5
68. Wu, X.M., Cai, J.F., Jiang, J.J., Zheng, D., Wei, Y.L., Zheng, W.S.: An economic framework for 6-DoF grasp detection. In: European Conference on Computer Vision (2024) 5
69. Wu, X.M., Fan, B., Liao, K., Jiang, J.J., Yang, R., Luo, Y., Wu, Z., Zheng, W.S., Loy, C.C.: VLANeXt: Recipes for building strong VLA models. In: International Conference on Machine Learning (2026) 5
70. Wu, X., DeTone, D., Frost, D.P., Shen, T., Xie, C., Yang, N., Engel, J.J., Newcombe, R.A., Zhao, H., Straub, J.: Sonata: Self-supervised learning of reliable point representations. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025) 13
71. Xie, C., Avetisyan, A., Howard-Jenkins, H., Siddiqui, Y., Straub, J., Newcombe, R.A., Balntas, V., Engel, J.J.: Human-in-the-loop local corrections of 3D scene layouts via infilling. arXiv preprint arXiv:2503.11806 (2025) 1
72. Yan, J., Gao, Y., Yang, Q., Wei, X., Xie, X., Wu, A., Zheng, W.: DreamView: Injecting view-specific text guidance into text-to-3d generation. In: European Conference on Computer Vision (2024) 4
73. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025) 13
74. Yang, J., Chen, X., Madaan, N., Iyengar, M., Qian, S., Fouhey, D.F., Chai, J.: 3D-GRAND: A million-scale dataset for 3D-llms with better grounding and less hallucination. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 29501–29512 (2025) 5
75. Ye, Q., Zeng, X., Li, F., Li, C., Fan, H.: Painting with Words: Elevating detailed image captioning with benchmark and alignment learning. In: International Conference on Learning Representations (2025) 6
76. Yeshwanth, C., Rozenberszki, D., Dai, A.: ExCap3D: Expressive 3D scene understanding via object captioning with varying detail. In: IEEE/CVF International Conference on Computer Vision (2025) 4, 6, 10
77. Yuan, Z., Yan, X., Liao, Y., Guo, Y., Li, G., Cui, S., Li, Z.: X-Trans2Cap: Cross-modal knowledge transfer using transformer for 3D dense captioning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022) 4
78. Zha, J., Fan, Y., Yang, X., Gao, C., Chen, X.: How to enable LLM with 3D capacity? A survey of spatial reasoning in LLM. In: International Joint Conference on Artificial Intelligence (2025) 4
79. Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., Huang, F., Zhou, J.: Qwen3 Embedding: Advancing text embedding

and reranking through foundation models. arXiv preprint arXiv:2506.05176 (2025) 8, 13
80. Zheng, D., Huang, S., Li, Y., Wang, L.: Learning from videos for 3D world: Enhancing MLLMs with 3D vision geometry priors. In: Neural Information Processing Systems (2025) 5
81. Zheng, D., Huang, S., Wang, L.: Video-3D LLM: learning position-aware video representation for 3D scene understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025) 5
82. Zheng, J., Zhang, J., Li, J., Tang, R., Gao, S., Zhou, Z.: Structured3D: A large photo-realistic dataset for structured 3D modeling. In: European Conference on Computer Vision (2020) 10
83. Zhou, J., Ma, T., Lin, K., Qiu, R., Wang, Z., Liang, J.: Mitigating the human-robot domain discrepancy in visual pre-training for robotic manipulation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025) 1
84. Zhou, J., Ye, K., Liu, J., Ma, T., Wang, Z., Qiu, R., Lin, K., Zhao, Z., Liang, J.: Exploring the limits of vision-language-action manipulations in cross-task generalization. In: Neural Information Processing Systems (2025) 1
85. Zhu, Z., Ma, X., Chen, Y., Deng, Z., Huang, S., Li, Q.: 3D-VisTA: Pre-trained transformer for 3D vision and text alignment. In: IEEE/CVF International Conference on Computer Vision (2023) 5
86. Zhu, Z., Zhang, Z., Ma, X., Niu, X., Chen, Y., Jia, B., Deng, Z., Huang, S., Li, Q.: Unifying 3D vision-language understanding via promptable queries. In: European Conference on Computer Vision (2024) 5