3D CoCa: Contrastive Learners are 3D Captioners

Ting Huang[1*] Zeyu Zhang[2*†] Yemin Wang[3*] Hao Tang[1‡]
[1]State Key Laboratory of Multimedia Information Processing,
School of Computer Science, Peking University
[2]The Australian National University [3]Xiamen University
[*]Equal contribution. [†]Project lead. [‡]Corresponding author: bjdxtanghao@gmail.com

Abstract

*3D captioning, which aims to describe the content of 3D scenes in natural language, remains highly challenging due to the inherent sparsity of point clouds and weak cross-modal alignment in existing methods. To address these challenges, we propose **3D CoCa**, a novel unified framework that seamlessly combines contrastive vision-language learning with 3D caption generation into a single architecture. We design a frozen CLIP vision-language backbone to provide rich semantic priors, a spatially-aware 3D scene encoder to capture geometric context, and a multi-modal decoder to generate descriptive captions. Unlike the prior two-stage methods that rely on explicit object proposals, 3D CoCa jointly optimizes contrastive and captioning objectives in a shared feature space, eliminating the need for external detectors or handcrafted proposals. This joint training paradigm yields stronger spatial reasoning and richer semantic grounding by aligning 3D and textual representations. Extensive experiments on the ScanRefer and Nr3D benchmarks demonstrate that 3D CoCa significantly outperforms current state-of-the-arts by 10.2% and 5.76% in CIDEr@0.5IoU, respectively. Code will be available at https://github.com/AIGeeksGroup/3DCoCa.*

1. Introduction

In recent years, 3D learning research has been increasing, driven by various practical applications such as robotics, autonomous driving, and augmented reality [13, 14, 21, 36]. Within this burgeoning field, the intersection of computer vision (CV) and natural language processing (NLP) has prompted researchers to strive to bridge the gap between visual perception and language expression, thus promoting the rise of cross-modal tasks such as visual captioning. The emergence of large-scale vision-language models has brought unprecedented breakthroughs in the generation of captions for 2D images. With the development of 3D vision-language datasets, 3D captions have also shown promising prospects. 3D captioning extends 2D image captioning and aims to accurately perceive the 3D structure of objects and generate reasonable descriptions by leveraging a comprehensive set of attribute details and contextual interaction information between objects and their surroundings. However, due to the sparsity of point clouds and the cluttered distribution of objects, describing objects within a 3D scene remains a particularly challenging endeavor.

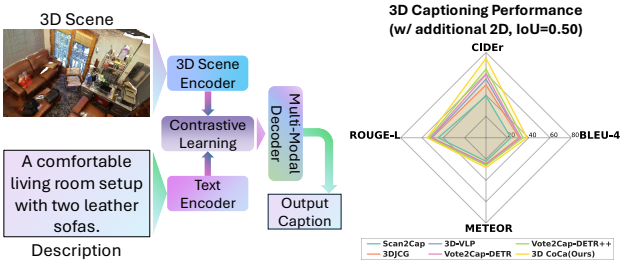

Figure 1. **Overview of the 3D CoCa framework and its captioning performance**. Left: The 3D CoCa model unifies contrastive learning and multimodal captioning in one framework. Right: Radar chart comparison of 3D CoCa and previous methods Scan2Cap [6], 3DJCG [3], 3D-VLP [40], Vote2Cap-DETR [11], Vote2Cap-DETR++ [12] on the ScanRefer [8] benchmark.

Early approaches to 3D captioning adopted a two-stage "detect-then-describe" paradigm, in which object proposals were first detected from point clouds and then described individually. For example, Scan2Cap [6] is the first attempt to integrate 3D object detection and caption generation into 3D scenes in a cascade manner. [18] introduced a novel 3D language pre-training approach that uses context-aware alignment and mutual masking to learn generic representations for 3D captioning tasks. Although effective, a two-stage pipeline can suffer from significant performance degradation. First, the detection stage usually produces redundant bounding boxes; thus, careful tuning using the Non-Maximum Suppression (NMS) [26] operation is required, which introduces additional hyperparameters and increases computational overhead. Second, the cascade design of the "detect-then-describe" process makes caption

generation highly dependent on the quality of the detection stage. In this context, the exploration of one-stage end-to-end 3D captioning models has attracted widespread attention. Vote2Cap-DETR [11] and its advanced version Vote2Cap-DETR++ [12] are notable examples that use the Transformer framework to simultaneously locate and describe objects during inference in a single forward pass, improving both efficiency and performance. Other recent approaches, such as BiCA [20], introduced a Bi-directional Contextual Attention mechanism to disentangle object localization from contextual feature aggregation in 3D scenes. The See-It-All (SIA) model [19] adopted a late aggregation strategy to capture both local object details and global contextual information with a novel aggregator. Moreover, TOD3Cap [17] employed a Bird's Eye View (BEV) representation for the generation of object proposals and integrated the Q-Former Relation with the LLaMA-Adapter to generate descriptive sentences, particularly for outdoor environments.

Despite progress, 3D captioning remains very challenging, especially in modeling spatial relations and aligning 3D visual data with textual semantics. Describing complex spatial arrangements requires the model to understand 3D geometry and relative object positions, which is non-trivial to encode and reason about. Bridging the gap between the 3D modality and language is also challenging. Existing methods treat vision and language as separate stages, with weak cross-modal interactions. This leads to suboptimal alignment between visual and textual representations.

These challenges point to the need for a unified framework that can enhance spatial reasoning and cross-modal alignment using strong visual-linguistic priors. Foundation models in vision-language research CoCa [37] have shown that contrastive pre-training on large image-text corpora yields representations with rich semantics and excellent alignment between modalities. Inspired by this, we hypothesize that incorporating such powerful priors into 3D captioning will significantly improve performance and generalization. This insight motivates us to design a 3D captioning approach that jointly learns spatially-grounded captions and visual-text alignments within a single end-to-end model, leveraging knowledge from large-scale vision-language training.

In this paper, we introduce 3D CoCa (Contrastive Captioner for 3D), as illustrated in Figure 1, a novel approach that integrates contrastive learning and caption generation into a unified model for 3D scenes. The core idea is to train a 3D scene encoder and a text encoder together with a shared contrastive learning objective, while simultaneously training a multi-modal decoder to generate captions. By coupling these tasks, 3D CoCa learns a joint feature space where 3D representations and captions are deeply aligned. The model leverages rich semantic knowledge from large-scale pre-training: we build on a vision-language backbone initialized with learned visual and linguistic features, injecting strong priors about objects and language into the 3D domain. This allows the model to recognize a wide range of concepts in the scene and associate them with the correct words. Furthermore, 3D CoCa is designed to be spatially aware-the 3D scene encoder preserves geometric structure, and the decoder's attention mechanism can attend to specific regions when wording the description. As a result, the generated captions capture not only object attributes, but also their precise spatial context, directly addressing the core difficulty of 3D captioning. In essence, our approach marries a powerful contrastive learner with a captioning model, demonstrating that contrastive learners are effective 3D captioners.

In summary, the main contributions of this work include:

- We present 3D CoCa, an end-to-end framework that combines contrastive vision-language learning with 3D captioning. It jointly learns to localize and describe objects from point clouds, removing reliance on external detectors.
- We design a 3D captioning framework that incorporates strong visual-linguistic priors from large-scale image-text pretraining. By introducing a contrastive alignment objective, the model enhances semantic grounding and cross-modal alignment, enabling more accurate and descriptive captions for complex 3D scenes.
- Extensive evaluations on benchmark datasets show that 3D CoCa achieves state-of-the-art captioning performance on Nr3D [1] (52.84% C@0.5) and Scanrefer [8] (77.13% C@0.5).

2. Related Work

3D captioning. 3D captioning involves localizing objects in a 3D scene and describing them in natural language. Early work like Scan2Cap [6] pioneered this task by leveraging point cloud data with spatial reasoning, marking a departure from conventional 3D detection pipelines focused only on classification and bounding boxes [4, 5, 41, 42]. Subsequent methods were built on this foundation with improved relational modeling. For example, the Multi-Order Relation Extraction (MORE) framework [16] introduced higher-order relationship reasoning, showing that richer spatial context leads to more informative and accurate captions.

The introduction of Transformer architectures has further accelerated progress in 3D captioning. SpaCap3D [33] employed a Transformer-based encoder–decoder with a spatially guided encoder to capture geometric context and an object-centric decoder for attribute-rich descriptions. χ-Trans2Cap [39] extended this idea by distilling knowledge from 2D vision-language models into a 3D captioner, effectively transferring semantic understanding from images to point clouds. Recent works strive for unified architec-

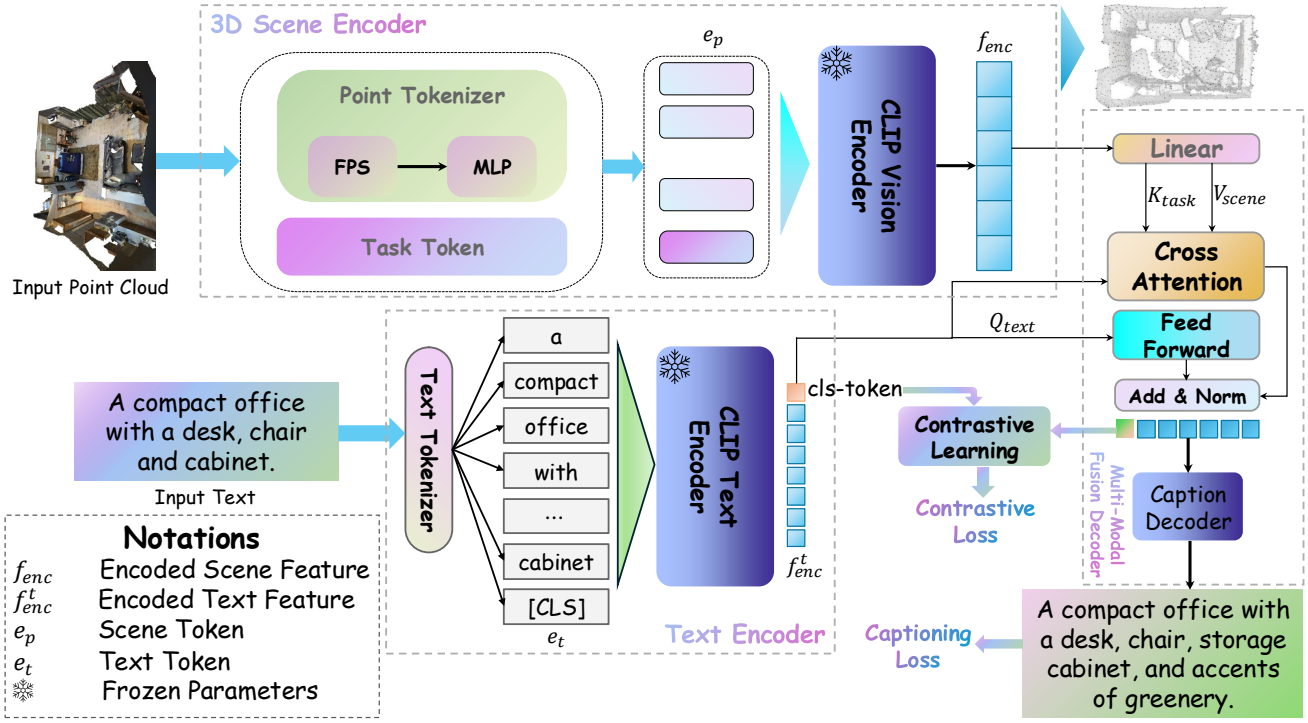

Figure 2. **Pipeline of 3D CoCa.** The input point cloud and textual description are processed by CLIP Vision and Text Encoders, respectively. Cross-attention mechanisms fuse these features within a Multi-Modal Decoder, enabling the generation of descriptive captions. The model training is guided by contrastive and captioning losses, promoting effective alignment between visual and textual modalities.

tures that handle multiple tasks: 3DJCG [3] uses shared Transformers to jointly optimize 3D captioning and visual grounding, and UniT3D [7] demonstrates that pre-training a Transformer on large-scale point cloud–text pairs can yield state-of-the-art results across diverse 3D scene understanding benchmarks.

Despite these advances, most approaches still follow a two-stage "detect-then-describe" paradigm [3, 6, 33, 39], where an object detector provides regions that are then described. This separation can cause error propagation and misalignment between the vision and language components. To overcome this limitation, end-to-end paradigms have been explored. Vote2Cap-DETR [11] and its improved variant Vote2Cap-DETR++ [12] reformulate dense captioning as a direct set-prediction task, similar to DETR in 2D vision. They jointly localize and caption objects in one stage, eliminating dependence on pre-trained detectors. Through a Transformer encoder–decoder with learnable queries and iterative refinement, these one-stage models achieve competitive performance while simplifying the pipeline.

3D pre-training and vision-language model. Another line of work has focused on pre-training 3D representations to provide stronger foundations for downstream tasks. Unsupervised 3D representation learning techniques can be categorized into global contrastive methods [25, 32] that learn holistic point cloud embeddings, local contrastive methods [34, 35] that distinguish fine-grained geometric structures or multi-view correspondences, and masked point modeling approaches [27, 38] that adapt masked autoencoding to 3D data. These approaches learn powerful geometric features; however, they operate purely on 3D geometry and lack grounding in natural language semantics.

To bridge this gap, researchers have explored 3D vision-language pre-training. For example, 3D-VLP [40] uses contrastive learning to align point cloud segments with text descriptions, producing representations that improve 3D captioning and visual grounding performance by injecting semantic knowledge. Similarly, UniT3D [7] showed that training on large-scale point cloud–caption pairs endows a unified model with strong multi-task 3D understanding capabilities. Such findings underscore the value of learning joint 3D language representations as a foundation for captioning.

3. The Proposed Method

3.1. Overview

In this section, we present **3D CoCa**, a unified framework designed to bridge the gap between 3D point cloud representation and natural language understanding for captioning. The framework integrates contrastive alignment with multimodal captioning, drawing inspiration from CLIP-style image-text models [30] and the Contrastive Captioner (CoCa) paradigm [37]. As shown in Figure 2, 3D CoCa is

composed of four main components: 3D Scene Encoder, Text Encoder, Contrastive Learning module, and Multi-Modal Fusion Decoder.

Unlike prior approaches that rely solely on either 2D images or 3D data, 3D CoCa transfers knowledge from large-scale 2D image-text pretraining and adapts it to the complexities of 3D point cloud understanding. To retain the strong visual and linguistic priors from CLIP, most of its pre-trained weights are kept frozen, with only lightweight 3D-specific modules introduced. The following subsections detail each component of the framework, followed by the joint training objectives that integrate them into a unified captioning model for 3D scenes.

3.2. 3D Scene Encoder

The role of the 3D scene encoder is to transform an unstructured point cloud into a set of latent tokens that capture the geometric and semantic content of the scene. The scene encoder integrates point-based 3D processing with a frozen 2D CLIP visual backbone to effectively capture geometric and semantic information. It comprises three components: (i) a point cloud tokenizer that partitions raw point clouds into patch tokens, (ii) a set of learnable task tokens that inject 3D captioning context, and (iii) a frozen CLIP Vision Transformer that encodes the concatenated token sequence. As shown in Figure 2 (top-left), the encoder transforms the unstructured 3D input into a structured representation suitable for multimodal reasoning.

Point cloud tokenizer. Given an input point cloud $P \in \mathbb{R}^{N\times(3+F)}$, where each of the N points is described by 3D coordinates (x, y, z) and F additional features (*e.g.,* color, normals, height, or multiview descriptors), we first convert the point cloud into a discrete token sequence suitable for transformer-based processing. To ensure spatial coverage, we use farthest point sampling (FPS) to select M representative points as patch centers. FPS promotes uniform sampling by reducing redundancy in dense regions while preserving structure in sparse areas. For each center, we gather its nearest neighbors K to construct a local patch, resulting in M patches $\{P_1, P_2, \ldots, P_M\}$, each containing K spatially adjacent points. Each patch is passed through a lightweight point-wise encoder, a series of multi-layer perceptrons (MLPs), to extract local geometric and appearance features. This yields a sequence of M point tokens, each represented as a D_p-dimensional embedding:

$$E_p(P) = [\mathbf{e}_{p_1}, \mathbf{e}_{p_2}, \ldots, \mathbf{e}_{p_M}] \in \mathbb{R}^{M\times D_p}, \tag{1}$$

where $\mathbf{e}_{p_i}$ denotes the embedding of the i-th patch. This tokenization strategy effectively transforms the unstructured 3D input into a structured sequence, balancing local detail (within each K-point patch) and global coverage (across the M sampled patches), enabling efficient downstream processing by the transformer backbone.

Task token. Although point tokens encode local visual features, they lack explicit task awareness. To guide the model toward 3D captioning, we introduce a set of learnable task tokens, embedding vectors that are prepended to the point token sequence. Inspired by prompt tuning [23], these task tokens are initialized with distinct fixed values (*e.g.,* indexed embeddings) and are updated during training. Performing as high-level prompts, the task tokens condition the model for the captioning objective. Through self-attention, they aggregate global semantic cues from the entire point cloud, such as scene layout and salient objects, which are critical for generating descriptive language. Effectively, they serve as a shared contextual anchor that biases the encoder toward language-relevant features in the 3D scene.

Frozen CLIP vision encoder. After obtaining the M point tokens and m_t task tokens, we concatenate them into a unified input sequence:

$$[\mathbf{e}_{p_1}, \ldots, \mathbf{e}_{p_M}; \mathbf{t}_1, \ldots, \mathbf{t}_{m_t}], \tag{2}$$

where $\mathbf{t}_j$ denotes the j-th task token embedding. This sequence of length $M + m_t$ is then fed into a frozen CLIP Vision Transformer [30], adapted from the original CLIP image encoder architecture. We keep all CLIP weights frozen to preserve their pre-trained visual representations and prevent catastrophic forgetting during training. This not only stabilizes optimization but also significantly reduces memory and computational overhead, as the majority of parameters remain fixed.

The CLIP encoder processes the token sequence and outputs a series of latent embeddings that jointly capture 3D geometry and task context. From these, we extract a global scene representation $f_{\text{enc}} \in \mathbb{R}^D$, which serves as a semantically rich embedding of the 3D scene conditioned on the captioning objective. This feature vector is used in the subsequent contrastive alignment with text representations.

3.3. Text Encoder

While the 3D scene encoder extracts visual features from point clouds, the text encoder transforms natural language descriptions into a semantically aligned embedding space. We adopt the Transformer-based CLIP text encoder [30] and keep its weights frozen to retain the rich linguistic knowledge acquired during large-scale pre-training. Freezing the text encoder ensures that all captions are assigned in the same semantic space as the CLIP visual representations, enabling effective alignment with the 3D scene embeddings.

Text tokenizer. Given an input sentence T, we tokenize it into a sequence of L subword tokens using a subword tokenizer, which ensures robustness to out-of-vocabulary words by decomposing them into known units. Each token w_i is assigned to an embedding D_t-dimensional vector via a

learned embedding table, yielding the text token sequence:

$$E_t(T) = [\mathbf{e}_{t_1}, \mathbf{e}_{t_2}, \ldots, \mathbf{e}_{t_L},] \in \mathbb{R}^{L \times D_t}, \quad (3)$$

where $\mathbf{e}_{t_i}$ denotes the embedding of the i-th token. A special beginning-of-sequence token is prepended to serve as a sentence-level aggregator. We further add positional encodings to preserve word order, which is essential for capturing the syntactic and semantic structures in natural language.

Frozen CLIP text encoder. The sequence of text embeddings $E_t(T)$ is processed by the CLIP text Transformer encoder, which consists of N_{te} layers of multi-head self-attention and feed-forward networks. Let $H^0 = E_t(T)$ be the input; the hidden states are updated as:

$$H^l = \text{TransformerBlock}^l(H^{l-1}), l \in [1, \ldots, N_{\text{te}}], \quad (4)$$

where each block includes self-attention, layer normalization, and MLP sublayers. All weights in the text encoder are kept frozen to preserve the linguistic knowledge acquired from large-scale image-text pretraining. This also helps prevent overfitting, especially given the limited size of 3D captioning datasets.

From the final layer, we extract the hidden state corresponding to the special [CLS] token as the global text representation: $f^t_{\text{enc}} \in \mathbb{R}^{D_t}$. This vector encodes the overall semantics of the caption and serves as the language-side embedding in the contrastive learning module. By leveraging a fixed CLIP text encoder, f^t_{enc} resides in the same embedding space as the CLIP visual features, enabling direct alignment with the 3D scene embedding f_{enc}.

3.4. Contrastive Learning Paradigm

To align the heterogeneous modalities of 3D point clouds and text, we employ a contrastive learning objective that maps both the 3D scene feature f_{enc} and the text feature f^t_{enc} into a shared embedding space. In this space, the matched 3D-text pairs are pulled closer, while mismatched pairs are pushed apart. This strategy mirrors the CLIP training paradigm, promoting cross-modal association through discriminative alignment. In the following, we detail the projection and normalization steps used to embed features in the shared space, followed by the formulation of the contrastive loss.

Feature alignment. Before computing similarity, we project the 3D scene feature f_{enc} and the text feature f^t_{enc} into a shared embedding space using learnable projection heads. Specifically, we apply two parallel two-layer MLPs:

$$\tilde{f}_{\text{enc}} = \text{MLP}_v\left(f_{\text{enc}}\right), \qquad \tilde{f}^t_{\text{enc}} = \text{MLP}_t\left(f^t_{\text{enc}}\right), \quad (5)$$

where MLP_v and MLP_t denote the projection heads for the 3D and text modalities, respectively. Each component consists of a linear layer, a ReLU activation, and a second linear layer. These modules not only unify the feature dimensionality, but also adapt the embeddings for optimal cross-modal alignment. To facilitate similarity computation, we apply L2 normalization to both projected vectors:

$$\hat{f}_{\text{enc}} = \frac{\tilde{f}_{\text{enc}}}{\left\|\tilde{f}_{\text{enc}}\right\|_2}, \qquad \hat{f}^t_{\text{enc}} = \frac{\tilde{f}^t_{\text{enc}}}{\left\|\tilde{f}^t_{\text{enc}}\right\|_2}. \quad (6)$$

This ensures that the features lie on the unit hypersphere, allowing their similarity to be computed as the cosine of the angle between them, which is an essential step for contrastive loss.

Contrastive loss function. With the projected and normalized features, we apply a contrastive learning objective to align paired 3D scenes and captions. Following the InfoNCE formulation popularized by CLIP, we consider a training batch of N 3D-text pairs and compute pairwise cosine similarities between all scene–caption combinations. For the i-th scene and the j-th text in the batch, the similarity is defined as:

$$\text{sim}\left(\hat{f}_{\text{enc},i}; \hat{f}^t_{\text{enc},j}\right) = \frac{\hat{f}_{\text{enc},i} \cdot \hat{f}^t_{\text{enc},j}}{\left\|\hat{f}_{\text{enc},i}\right\| \left\|\hat{f}^t_{\text{enc},j}\right\|}, \quad (7)$$

which corresponds to the dot product of unit-normalized vectors. The contrastive loss maximizes the similarity between matching pairs ($i = j$), while minimizing it for mismatched pairs ($i \neq j$). For each scene i, the loss is given by:

$$\mathcal{L}_{\text{Con}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp\left(\text{sim}\left(\hat{f}_{\text{enc},i}; \hat{f}^t_{\text{enc},i}\right)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}\left(\hat{f}_{\text{enc},i}; \hat{f}^t_{\text{enc},j}\right)/\tau\right)}, \quad (8)$$

where τ is a learnable temperature parameter that controls the sharpness of the softmax distribution. This loss encourages the model to assign high similarity to correct 3D-text pairs while suppressing alignment with incorrect matches. By training with this objective, the model learns a shared embedding space in which 3D scenes and their corresponding captions are closely aligned, facilitating effective cross-modal understanding.

3.5. Multi-Modal Fusion Decoder

The last component of 3D CoCa is the multi-modal fusion decoder, which is responsible for generating natural language descriptions conditioned on the input 3D scene. It is implemented as an autoregressive Transformer decoder equipped with cross-attention mechanisms that integrate 3D scene context at each generation step.

Performing as a conditional language model, the decoder generates captions token by token. It begins with a special start-of-sequence token and, at each time step t, conditions on previously generated tokens $y_{<t}$ through causal

self-attention to model intra-sentence dependencies. Simultaneously, it attends to the 3D scene embedding via cross-attention, enabling it to incorporate visual context into each prediction: $P(y_t \mid y_{<t}, f_{\text{enc}})$. By leveraging the semantically aligned 3D features obtained from the contrastive stage, the decoder produces descriptions that are not only grammatically coherent but also faithful to the 3D content. The integration of causal and cross-modal attention ensures that the output is linguistically fluent and visually grounded.

Cross-Attention mechanism. To integrate 3D visual context into the captioning process, the decoder incorporates cross-modal attention layers. In each decoder block, a cross-attention module enables the decoder to attend to the encoded 3D scene tokens produced by the scene encoder (Section 3.2). Formally, let Q_{text} denote the query matrix from the decoder's current hidden states, and K_{task}, V_{scene} be the key and value matrices derived from the 3D scene embeddings. The cross-attention is computed as:

$$\text{Attention}\left(Q_{\text{text}}, K_{\text{task}}, V_{\text{scene}}\right) = \text{softmax}\left(\frac{Q_{\text{text}} K_{\text{task}}^{\top}}{\sqrt{d_k}}\right) V_{\text{scene}}, \quad (9)$$

where d_k is the key dimensionality. This operation enables each position in the decoder to selectively incorporate relevant 3D information, guided by the attention weights. By conditioning the decoder on the scene features at every generation step, the cross-attention mechanism ensures that the output captures both global context and fine-grained spatial details from the 3D input. The resulting captions are grounded in the visual content and reflect the structural layout of the scene.

To support both linguistic coherence and visual grounding, the cross-attention layers are interleaved with self-attention layers within the decoder. This design allows for a dynamic interplay between textual dependencies and visual context, enabling the model to generate fluent and semantically accurate descriptions.

Training objectives and joint optimization. We train the multi-modal decoder using a combination of contrastive loss and captioning loss, jointly optimizing both to enforce cross-modal alignment and improve generation quality. The contrastive loss $\mathcal{L}_{\text{Con}}$ (Eq. (8)) is applied at the encoder level to align 3D and text features in a shared embedding space. This alignment provides a strong initialization for the decoder and guides its cross-attention to focus on semantically relevant visual regions.

The decoder itself is supervised with a standard cross-entropy loss over the captioning task. Given a predicted caption $\hat{Y} = (\hat{y}_1, \ldots, \hat{y}_L)$ and the corresponding ground-truth sequence $Y = (y_1, \ldots, y_L)$, the captioning loss is defined as:

$$\mathcal{L}_{\text{Cap}} = -\sum_{t=1}^{L} \log P\left(\hat{y}_t = y_t \mid \hat{y}_{<t}, f_{\text{enc}}\right), \quad (10)$$

Algorithm 1: 3D CoCa Algorithm

Require: Point cloud data P, Text input T
Ensure: Generated caption $\hat{C}$

Point Cloud & Text Input Processing:
$\mathbf{E}_p \leftarrow$ Point cloud tokenizer(P) {Tokenize input point cloud into sequence}
$\mathbf{E}_t \leftarrow$ Text tokenizer(T) {Tokenize input text into sequence}
Feature Encoding via Frozen CLIP Encoders:
$\mathbf{f}_{\text{enc}} \leftarrow \text{CLIP}_{\text{visual}}(\mathbf{E}_p)$ {Frozen CLIP visual encoder}
$\mathbf{f}_{\text{enc}}^{t} \leftarrow \text{CLIP}_{\text{text}}(\mathbf{E}_t)$ {Frozen CLIP text encoder}
Feature Alignment & Contrastive Learning:
$(\hat{\mathbf{f}}_{\text{enc}}, \hat{\mathbf{f}}_{\text{enc}}^{t}) \leftarrow$ Feature alignment & Normalize$\left(\mathbf{f}_{\text{enc}}, \mathbf{f}_{\text{enc}}^{t}\right)$
$\mathcal{L}_{\text{Con}} \leftarrow \text{InfoNCE}\left(\hat{\mathbf{f}}_{\text{enc}}, \hat{\mathbf{f}}_{\text{enc}}^{t}\right)$ {Contrastive loss for matching vs. non-matching pairs}
Update alignment layers using $\mathcal{L}_{\text{Con}}$
Multi-modal Decoding & Caption Generation:
$\hat{C} \leftarrow \text{TransformerDecoder}(\mathbf{f}_{\text{enc}})$ {Cross-attention over $\mathbf{f}_{\text{enc}}$, autoregressive generation}
Joint Optimization Objective:
$\mathcal{L}_{\text{Cap}} \leftarrow \text{CrossEntropy}\left(\hat{C}, C_{gt}\right)$ {Caption generation loss}
$\mathcal{L}_{\text{Total}} \leftarrow \mathcal{L}_{\text{Cap}} + \lambda \cdot \mathcal{L}_{\text{Con}}$

where f_{enc} is the global 3D scene embedding used to condition the decoder via cross-attention.

The overall training objective combines both losses:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{Con}} + \lambda \cdot \mathcal{L}_{\text{Cap}}, \quad (11)$$

where λ is a weighting hyperparameter that controls the trade-off between alignment and generation. In practice, we set λ to assign approximately equal importance to both objectives.

This joint training scheme creates a mutually reinforcing learning signal: contrastive alignment ensures that the encoder produces scene features that are accessible and informative for the decoder, while caption supervision encourages the encoder to capture fine-grained, text-relevant details. As a result, 3D CoCa learns to generate captions that are not only fluent and descriptive but also precisely grounded in the 3D scene content, as summarized in algorithm 1.

4. Experiments

4.1. Datasets and Evaluation Metrics

Datasets. We evaluate 3D captioning performance on two widely used benchmarks: ScanRefer [8] and Nr3D [1], both of which contain human-annotated descriptions of 3D objects and scenes. ScanRefer comprises 36,665 descriptions for 7,875 objects across 562 scenes, while Nr3D includes 32,919 descriptions for 4,664 objects in 511 scenes. Both datasets are derived from ScanNet [15], which provides a large-scale collection of 1,201 indoor 3D scenes. For evaluation, we use the standard validation splits: ScanRefer con-

Table 1. **Comparison of various methods on the ScanRefer dataset [8].** We evaluate the performance of each method, with and without additional 2D input, at IoU thresholds of 0.25 and 0.5. Metrics include CIDEr (C) [31], BLEU-4 (B-4) [28], METEOR (M) [2], and ROUGE-L (R) [22]. Our proposed 3D CoCa achieves state-of-the-art results across all settings.

Method	w/o additional 2D input								w/ additional 2D input							
	IoU = 0.25				IoU = 0.50				IoU = 0.25				IoU = 0.50			
	C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑	C↑	B-4↑	M↑	R↑
Scan2Cap [6]	53.73	34.25	26.14	54.95	35.20	22.36	21.44	43.57	56.82	34.18	26.29	55.27	39.08	23.32	21.97	44.78
MORE [16]	58.89	35.41	26.36	55.41	38.98	23.01	21.65	44.33	62.91	36.25	26.75	56.33	40.94	22.93	21.66	44.42
SpaCap3d [33]	58.06	35.30	26.16	55.03	42.76	25.38	22.84	45.66	63.30	36.46	26.71	55.71	44.02	25.26	22.33	45.36
3DJCG [3]	60.86	39.67	27.45	59.02	47.68	31.53	24.28	51.80	64.70	40.17	27.66	59.23	49.48	31.03	24.22	50.80
D3Net [9]	-	-	-	-	-	-	-	-	-	-	-	-	46.07	30.29	24.35	51.67
3D-VLP [40]	64.09	39.84	27.65	58.78	50.02	31.87	24.53	51.17	70.73	41.03	28.14	59.72	54.94	32.31	24.83	51.51
Vote2Cap-DETR [11]	71.45	39.34	28.25	59.33	61.81	34.46	26.22	54.40	72.79	39.17	28.06	59.23	59.32	32.42	25.28	52.53
Unit3D [7]	-	-	-	-	-	-	-	-	-	-	-	-	46.69	27.22	21.91	45.98
Vote2Cap-DETR++ [12]	76.36	41.37	28.70	60.00	67.58	37.05	26.89	55.64	77.03	40.99	28.53	59.59	64.32	34.73	26.04	53.67
BiCA [20]	78.42	41.46	28.82	60.02	68.46	38.23	27.56	58.56	78.35	41.20	28.82	59.80	66.47	36.13	26.71	54.54
See-It-All [19]	78.68	43.25	29.21	**63.06**	73.22	40.91	28.19	**60.46**	78.05	42.16	28.74	**61.70**	69.86	37.89	27.04	**57.33**
3D CoCa (Ours)	**85.42**	**45.56**	**30.95**	61.98	**77.13**	**41.23**	**28.52**	57.40	**86.12**	**44.79**	**30.75**	61.45	**74.52**	**38.42**	**28.03**	55.23

Table 2. **Comparison on Nr3D [1] at IoU=0.5.** Our model outperforms existing methods, demonstrating higher CIDEr (C) [31], BLEU-4 (B-4) [28], METEOR (M) [2], and ROUGE-L (R) [22] scores.

Method	C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑
Scan2Cap [6]	27.47	17.24	21.80	49.06
SpaCap3d [33]	33.71	19.92	22.61	50.50
D3Net [9]	33.85	20.70	23.13	53.38
3DJCG [3]	38.06	22.82	23.77	52.99
Vote2Cap-DETR [11]	43.84	26.68	25.41	54.43
Vote2Cap-DETR++ [12]	47.08	27.70	25.44	55.22
BiCA [20]	48.77	28.35	**25.60**	55.81
3D CoCa (Ours)	**52.84**	**29.29**	25.55	**56.43**

tains 9,508 descriptions for 2,068 objects in 141 scenes, while Nr3D contains 8,584 descriptions for 1,214 objects in 130 scenes. All evaluation scenes are drawn from the 312 3D scenes in the ScanNet validation set. This setup ensures a consistent evaluation protocol across both datasets.

Evaluation metrics. We evaluate model performance using four standard metrics: CIDEr [31], BLEU-4 [28], METEOR [2], and ROUGE-L [22], denoted C, B-4, M, and R, respectively. Following prior works [3, 6, 11, 16, 33], we apply Non-Maximum Suppression (NMS) to filter duplicate object proposals before evaluation. To assess the quality of generated captions while accounting for localization accuracy, we adopt the $m@k$IoU metric [6]. Given a total of N annotated objects, it is defined as:

$$m@k\text{IOU} = \frac{1}{N}\sum_{i=1}^{N} m\left(\hat{c}_i, C_i\right) \cdot \mathbb{I}\left\{\text{IoU}\left(\hat{b}_i, b_i\right) \geq k\right\}, \quad (12)$$

where $\hat{c}_i$ and C_i denote the predicted and ground-truth captions, $\hat{b}_i$ and b_i are the predicted and ground-truth 3D bounding boxes, and $m(\cdot)$ is any captioning metric (*e.g.,* CIDEr, METEOR, BLEU-4, ROUGE-L).

4.2. Implementation Details

We detail the implementation settings for our model and the baselines. The input point cloud $\mathcal{P} \in \mathbb{R}^{40,000\times 10}$ consists of 40,000 points, each represented by its absolute 3D position and additional per-point features. In the "w/o additional 2D" setting, these features include color, normal, and height, while in the "w/ additional 2D" setting, the color channel is replaced by 128-dimensional multiview features extracted from 2D images using ENet [10], following the protocol in [6].

We train all models for 1,080 epochs on ScanRefer [8] and Nr3D [1] using a combination of cross-entropy loss and contrastive loss. Optimization is performed using AdamW [24] with a learning rate of 0.1, a batch size of 4, and a cosine annealing learning rate schedule. All experiments are conducted on a single NVIDIA RTX 4090 GPU.

4.3. Comparative Study

We evaluate our method on ScanRefer and Nr3D using four standard captioning metrics: CIDEr (C) [31], METEOR (M) [2], BLEU-4 (B-4) [28], and ROUGE-L (R) [22]. Evaluation follows the IoU thresholds of 0.25 and 0.5 for ScanRefer (Table 1) and 0.5 for Nr3D (Table 2); the dashed ("-") indicates that results are not reported in the original or follow-up works.

Scanrefer. As shown in Table 1, our model consistently outperforms existing approaches across all settings and IoU thresholds, demonstrating stronger spatial understanding and linguistic expressiveness.

Nr3D. On Nr3D (Table 2), 3D CoCa achieves notable gains in CIDEr and BLEU-4, reflecting improved reasoning over natural, free-form descriptions.

Qualitative results. Figure 3 provides qualitative comparisons on ScanRefer, where our method generates more fluent and semantically grounded captions than Vote2Cap-DETR++ [12], aligning better with ground-truth descriptions.

4.4. Ablation Study

Contrastive learning loss impact analysis. We first study the effect of incorporating contrastive learning and analyze

Table 3. **The impact of contrastive learning loss weight λ on ScanRefer.** When λ=1, the performance is the best, indicating the role of contrastive learning.

λ	C@0.25↑	B-4@0.25↑	M@0.25↑	R@0.25↑
0.0	74.12	40.98	27.45	58.76
0.1	77.30	41.80	28.10	59.60
0.5	79.55	42.55	28.75	60.40
1.0	**85.42**	**45.56**	**30.95**	**61.98**
2.0	76.89	41.50	28.00	59.30

Table 4. **The impact of different caption generation decoders.** Comparison of the description indicators of the original GPT-2 generator and the CoCa-style multimodal decoder in this paper under the same visual features.

Caption Decoder	C@0.25↑	B-4@0.25↑	M@0.25↑	R@0.25↑
GPT-2 Captioner (Baseline)	76.20	41.00	27.80	59.50
CoCa Transformer (Ours)	**85.42**	**45.56**	**30.95**	**61.98**

Table 5. **Comparison of the impact of different 3D point cloud encoder architectures on description performance.**

Encoder Architecture	C@0.25↑	B-4@0.25↑	M@0.25↑	R@0.25↑
PointNet++ (Baseline)	72.48	38.95	26.80	56.30
3D CoCa (Ours)	**85.42**	**45.56**	**30.95**	**61.98**

the model sensitivity to different values of the weighting coefficient λ. Specifically, we vary $\lambda \in \{0, 0.1, 0.5, 1.0, 2.0\}$ to control the strength of the contrastive objective. As shown in Table 3, the model without contrastive loss ($\lambda = 0$) yields the weakest performance across all metrics. Introducing contrastive supervision significantly improves results: for instance, CIDEr increases from 74.12 to 79.55 as λ increases from 0 to 0.5, with the best overall performance achieved at λ = 1.0. When the weight increases to $\lambda = 2.0$, the performance decreases slightly, although it still surpasses the baseline without contrastive learning. These results suggest that a moderate contrastive objective enhances cross-modal alignment and semantic grounding, thus improving the quality of 3D scene descriptions. Overemphasizing the contrastive loss, however, may hinder generation fluency due to overly constrained representations.

Decoder architecture comparison. Then, we evaluate the impact of the captioning decoder architecture while keeping the visual encoder output fixed. As shown in Table 4, this substitution leads to a notable drop in captioning performance across all metrics. The results highlight the effectiveness of the CoCa-style decoder in integrating contrastively aligned visual features during generation. Its cross-attentional design enables tighter coupling between visual context and textual output, producing descriptions that are more semantically informative and better grounded in the 3D scene content.

Point cloud encoder analysis. Finally, we compare the proposed EPCL-based point cloud encoder against the conventional PointNet++ [29] encoder under identical settings. As shown in Table 5, our EPCL-based encoder yields consistently superior performance. These results demonstrate that the EPCL framework, when combined with pre-trained CLIP features, significantly enhances both the semantic representation and spatial modeling of 3D scenes. The richer and more structured scene understanding enables the model to generate captions that are more accurate, detailed, and better grounded in the visual context.

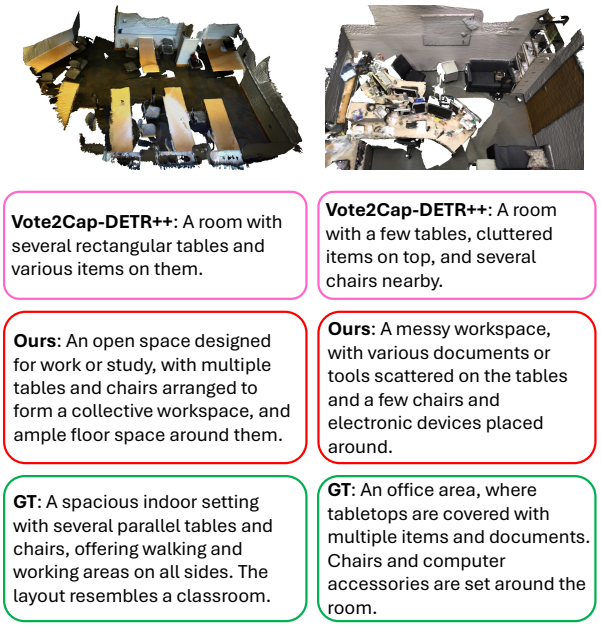

Figure 3. **Qualitative results on ScanRefer [8].** A visual comparison on the ScanRefer [8] dataset showcasing indoor scenes described by Vote2Cap-DETR++ [12], our method (Ours), and the ground truth (GT), highlighting differences in descriptive accuracy and style.

5. Conclusion

In this work, we propose 3D CoCa, a unified contrastive-captioning framework for 3D vision-language tasks. By jointly learning contrastive 3D-text representations and caption generation within a single model, 3D CoCa eliminates the need for any explicit 3D object detectors or proposal stages. This unified approach enables direct 3D-to-text alignment in a shared feature space, leading to improved spatial reasoning and more precise semantic grounding compared to previous methods. Experiments on two widely used datasets validate that our proposed 3D CoCa model significantly outperforms existing methods across standard captioning metrics and demonstrates the benefits of our contrastive learning strategy. We believe that 3D CoCa establishes a strong foundation for future research on unified 3D vision-language models and open-world spatial understanding.

Acknowledgements. This work was supported by the Fundamental Research Funds for the Central Universities, Peking University.

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. *16th European Conference on Computer Vision (ECCV)*, 2020. 2, 6, 7

[2] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, 2005. Association for Computational Linguistics. 7

[3] Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16443–16452, 2022. 1, 3, 7

[4] Guohui Cai, Ying Cai, Zeyu Zhang, Yuanzhouhan Cao, Lin Wu, Daji Ergu, Zhinbin Liao, and Yang Zhao. Medical ai for early detection of lung cancer: A survey. *arXiv preprint arXiv:2410.14769*, 2024. 2

[5] Guohui Cai, Ruicheng Zhang, Hongyang He, Zeyu Zhang, Daji Ergu, Yuanzhouhan Cao, Jinman Zhao, Binbin Hu, Zhinbin Liao, Yang Zhao, et al. Msdet: Receptive field enhanced multiscale detection for tiny pulmonary nodule. *arXiv preprint arXiv:2409.14028*, 2024. 2

[6] DaveZhenyu Chen, Ali Gholami, Matthias Niesner, and AngelX. Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 3, 7

[7] DaveZhenyu Chen, Ronghang Hu, Xinlei Chen, Matthias Nießner, and AngelX. Chang. Unit3d: A unified transformer for 3d dense captioning and visual grounding. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 18063–18073, 2023. 3, 7

[8] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. *16th European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 6, 7, 8

[9] Dave Zhenyu Chen, Qirui Wu, Matthias Nießner, and Angel X Chang. D3net: A speaker-listener architecture for semi-supervised dense captioning and visual grounding in rgb-d scans. *arXiv preprint arXiv:2112.01551*, 2021. 7

[10] Jintai Chen, Biwen Lei, Qingyu Song, Haochao Ying, Danny Z. Chen, and Jian Wu. A hierarchical graph network for 3d object detection on point clouds. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 389–398, 2020. 7

[11] Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Gang Yu, and Tao Chen. End-to-end 3d dense captioning with vote2cap-detr. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 11124–11133, 2023. 1, 2, 3, 7

[12] Sijin Chen, Hongyuan Zhu, Mingsheng Li, Xin Chen, Peng Guo, Yinjie Lei, Gang Yu, Taihao Li, and Tao Chen. Vote2cap-detr++: Decoupling localization and describing for end-to-end 3d dense captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(11): 7331–7347, 2024. 1, 2, 3, 7, 8

[13] Xin Chen, Anqi Pang, Yang Wei, Wang Peihao, Lan Xu, and Jingyi Yu. Tightcap: 3d human shape capture with clothing tightness field. *ACM Transactions on Graphics (Presented at ACM SIGGRAPH)*, 2021. 1

[14] Xin Chen, Anqi Pang, Wei Yang, Yuexin Ma, Lan Xu, and Jingyi Yu. Sportscap: Monocular 3d human motion capture and fine-grained understanding in challenging sports videos. *International Journal of Computer Vision*, page 2846–2864, 2021. 1

[15] 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. 6

[16] Yang Jiao, Shaoxiang Chen, Zequn Jie, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. More: Multi-order relation mining for dense captioning in 3d scenes. In *In Proceedings of the European conference on computer vision*, page 528–545, 2022. 2, 7

[17] Bu Jin, Yupeng Zheng, Pengfei Li, Weize Li, Yuhang Zheng, Sujie Hu, Xinyu Liu, Jinwei Zhu, Zhijie Yan, Haiyang Sun, Kun Zhan, Peng Jia, Xiaoxiao Long, Yilun Chen, and Hao Zhao. Tod3cap: Towards 3d dense captioning in outdoor scenes. In *In Proceedings of the European conference on computer vision*, page 367–384, 2025. 2

[18] Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, and Yinjie Lei. Context-aware alignment and mutual masking for 3d-language pre-training. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 10984–10994, 2023. 1

[19] Minjung Kim, Hyung Lim, SeungHwan Kim, Soonyoung Lee, Bumsoo Kim, and Gunhee Kim. See it all: Contextualized late aggregation for 3d dense captioning. In *Findings of the Association for Computational Linguistics ACL 2024*, page 3395–3405, 2024. 2, 7

[20] Minjung Kim, HyungSuk Lim, Soonyoung Lee, Bumsoo Kim, and Gunhee Kim. Bi-directional contextual attention for 3d dense captioning. In *In Proceedings of the European conference on computer vision*, page 385–401, 2025. 2, 7

[21] Yongbin Liao, Hongyuan Zhu, Yanggang Zhang, Chuangguan Ye, Tao Chen, and Jianchao Fan. Point cloud instance segmentation with semi-supervised bounding-box mining. *Cornell University - arXiv,Cornell University - arXiv*, 2021. 1

[22] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. 7

[23] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages

61–68, Dublin, Ireland, 2022. Association for Computational Linguistics. 4

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 7

[25] Guofeng Mei, Xiaoshui Huang, Juan Liu, Jian Zhang, and Qiang Wu. Unsupervised point cloud pre-training via contrasting and clustering. In *2022 IEEE International Conference on Image Processing (ICIP)*, 2022. 3

[26] A. Neubeck and L. Van Gool. Efficient non-maximum suppression. In *18th International Conference on Pattern Recognition (ICPR'06)*, pages 850–855, 2006. 1

[27] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part II*, pages 604–621. Springer, 2022. 3

[28] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, page 311–318, USA, 2002. Association for Computational Linguistics. 7

[29] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 8

[30] 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, 2021. 3, 4

[31] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015. 7

[32] Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matt J. Kusner. Unsupervised point cloud pre-training via occlusion completion. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[33] Heng Wang, Chaoyi Zhang, Jianhui Yu, and Weidong Cai. Spatiality-guided transformer for 3d dense captioning on point clouds. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, page 1393–1400, 2022. 2, 3, 7

[34] Ziyi Wang, Xumin Yu, Yongming Rao, Jie Zhou, and Jiwen Lu. Take-a-photo: 3d-to-2d generative pre-training of point cloud models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5617–5627, 2023. 3

[35] Saining Xie, Jiatao Gu, Demi Guo, Charles R. Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Computer Vision – ECCV 2020*, pages 574–591, Cham, 2020. Springer International Publishing. 3

[36] Fukun Yin, Zilong Huang, Tao Chen, Guozhong Luo, Gang Yu, and Bin Fu. Dcnet: Large-scale point cloud semantic segmentation with discriminative and efficient feature aggregation. *IEEE Transactions on Circuits and Systems for Video Technology*, page 1–1, 2023. 1

[37] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022. 2, 3

[38] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[39] Zhihao Yuan, Xu Yan, Yinghong Liao, Yao Guo, Guanbin Li, Shuguang Cui, and Zhen Li. X -trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 8553–8563, 2022. 2, 3

[40] Taolin Zhang, Sunan He, Tao Dai, Zhi Wang, Bin Chen, and Shu-Tao Xia. Vision-language pre-training with object contrastive learning for 3d scene understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38 (7):7296–7304, 2024. 1, 3, 7

[41] Zeyu Zhang, Nengmin Yi, Shengbo Tan, Ying Cai, Yi Yang, Lei Xu, Qingtai Li, Zhang Yi, Daji Ergu, and Yang Zhao. Meddet: Generative adversarial distillation for efficient cervical disc herniation detection. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 4024–4027. IEEE, 2024. 2

[42] Rui Zhao, Zeyu Zhang, Yi Xu, Yi Yao, Yan Huang, Wenxin Zhang, Zirui Song, Xiuying Chen, and Yang Zhao. Peddet: Adaptive spectral optimization for multimodal pedestrian detection. *arXiv preprint arXiv:2502.14063*, 2025. 2