A Unified Multi-Modal Approach for 3D Referring Expression Segmentation

KESHEN ZHOU

M.Phil

THE UNIVERSITY OF
SYDNEY

Supervisor: Dr. Tongliang Liu
Associate Supervisor: Dr. Baosheng Yu

A thesis submitted in fulfilment of
the requirements for the degree of
Master of Philosophy

School of Computer Science
Faculty of Engineering
The University of Sydney
Australia

30 August 2025

Statement of Originality

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Student Name: Keshen Zhou **Date:** 30 August 2025

Student Signature:

Authorship Attribution Statement

I declare that the thesis is my own original work and has not been previously published. None of the chapters in this thesis are published as papers or edit book chapter. I confirm that this thesis does not contain any material previously published or written by myself or others, except where due reference is made in the text of the thesis. All sources of information and assistance have been specifically acknowledged.

Student Name: Keshen Zhou **Date:** 30 August 2025

Student Signature:

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Supervisor Name: Tongliang Liu **Date:** 30 August 2025

Supervisor Signature:

Statement of Use of Gen AI

During the preparation of this thesis, ChatGPT from OpenAI was used for text enhancement purposes, including logic corrections and grammatical corrections.

For example, I modified the sentence "The model need be training with more datas, which may increases the computation cost and memory usage." to "The model needs to be trained with more data, which may increase computational cost and memory usage" for grammatical correctness. For all the generated results I received from Chatgpt, I have thoroughly reviewed all of them which are used to correct any errors, inaccuracies, or biases, and made adjustments as needed. I take full responsibility for the submitted thesis, ensure the work is my own, and have used generative AI within acceptable parameters.

Student Name: Keshen Zhou **Date:** 30 August 2025

Student Signature:

Abstract

Generalised 3D Referring Expression Segmentation (3D-GRES) segments exact 3D objects described by free-form language, even when descriptions match multiple targets, single targets, or zero targets. Most existing methods rely solely on sparse, colour-poor point clouds, neglecting the complementary semantics richness of multi-view RGB images. In this paper, we propose IS-RES, a unified multi-modal framework that integrates RGB images and point clouds for 3D-GRES. Specifically, IS-RES extracts the instance mask by Segment Anything Model(SAM), obtains both dense and instance-aware 2D embeddings through CLIP, and unprojects 2D embeddings into 3D point clouds via confidence-weighted pixel-to-point association. A progressive multi-level fusion strategy is applied to transform fragmented multi-modal features into hierarchical representations, enabling adaptive alignment between instance-level semantics and geometric structures. Extensive experiments demonstrate that IS-RES achieves state-of-the-art performance on both ScanRefer and Multi3DRefer benchmarks, with significant improvements in challenging scenarios involving multiple instances and complex spatial relationships.

Acknowledgements

As time passes, my postgraduate journey is coming to a close. Looking back over the past two years, I have experienced moments of frustration and even times when I wanted to give up, yet what lingers most is a profound sense of gratitude. This was my first real experience with research. There were lots of moments of frustration: moments when the code would not run, when experiments failed, when self-doubt crept in. Yet, more than anything, I am grateful: Stepping into research for the first time has marked a real milestone in my life and pushed me to grow, not just in knowledge but also in mindset and mentally.

I've always had a habit of keeping a diary, and sometimes when I look back at my early entries, I realize just how much I've changed—what I used to care about, what I used to think about every day—it's all evolved.

First of all, I would like to express my deepest respect and endless gratitude to Professor Tongliang Liu. Throughout my MPhil studies, Professor Liu has truly been my role model, showing me what it means to approach research with his rigorous academic standards, deep knowledge, and sharp insight. Watching him work has taught me that good research isn't just about having ideas—it's about staying focused, working hard, and never settling for anything less than your best.

Within the TML Group, formed and led by our professor, I met and learned from distinguished researchers in both academia and industry, opportunities I could never have imagined as an undergraduate. Without his generous support and patient guidance, completing this journey would have been unimaginable. Looking ahead, I hope to find my own passion and pursue it with the same dedication and commitment I've witnessed in Professor Liu.

I am also deeply grateful to my co-supervisor, Dr. Runnan Chen. His guidance opened the door to the world of research for me, while also gaining a treasured friendship. Throughout this journey, Dr. Chen was always there with thoughtful advice and careful guidance whenever I needed it. His insights have been invaluable, and his help with everything from developing

research ideas to designing experiments and writing papers was absolutely indispensable. His patience and rigorous approach helped me work through difficulties and find solutions I couldn't have reached on my own.

My deepest thanks also go to my parents, themselves professors, whose unwavering support in daily life has been an inexhaustible source of strength. Sharing my research ideas with them often yielded fresh perspectives, and their encouragement has fueled my pursuit of excellence. Sometimes feedback could be stressful, but only those who care you most will tell you what you need to hear in the most direct and honest way. I am so lucky to be born into such a warm and loving family.

I am grateful as well to my friends and all of our teammates in TML Lab, in particular, Mr. Aoran Liu, Mr. Zhaoqing Wang, Mr. Yaxuan Song, Mr. Xiangyu Sun, Mr. Suqin Yuan, Mr. Yexiong Lin, Mr. Yingzhen Wang, Mr. Chaojian Yu and Mr. Li He and many others too numerous to list. Whenever setbacks arose, you offered guidance and motivation helps me.

Completion of this thesis does not mark the end of learning; life is about learning, discovering and exploring new possibilities. While this work is the result of my own efforts, it is equally the culmination of the dedication and contributions of all who have supported me along the way.

Once again, I offer my sincerest gratitude to everyone who has helped and encouraged me. Life is just beginning, and the road ahead is long. With a grateful heart, I will continue striving for excellence, determined to live up to the trust and support you have placed in me.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

3D Referring Expression Segmentation (3D-RES) aims to segment specific objects within 3D scenes based on natural language descriptions, expanding beyond visual grounding task [3, 5, 93, 1, 8, 47] to provide precise mask-level understanding [27, 42, 73, 72, 6]. Given a 3D scene and free-form textual expressions such as "the blue mug on the corner of the desk," the task requires bridging the semantic gap between linguistic descriptions and spatial-geometric representations to produce accurate 3D segmentation masks, which are for enabling robotic manipulation, augmented reality, and human-computer interaction.

Early 3D-RES methods adopted two-stage frameworks [27], which proved computationally expensive and suboptimal [73], motivating the development of end-to-end architectures. Recent one-stage approaches [73, 57, 21, 72, 39, 78] employ query-based mechanisms that directly predict segmentation masks from unified scene representations. Notably, 3D-STMN [73] introduced superpoints [63, 34] as geometric priors, establishing a framework widely adopted by subsequent works [72, 26, 6], with 3D-GRES [72] further extending the task to generalized settings [93, 41] that address scenarios where referring expressions may correspond to multiple objects or no objects.

Despite these advances, these methods predominantly rely on 3D point cloud data alone, which inherently suffers from sparsity and lacks rich texture information crucial for interpreting complex visual attributes described in natural language. While point clouds captured by depth-aware sensors such as LiDAR encode precise geometry but remain sparsely distributed and largely colour-blind, multi-view RGB images provide dense texture information that is essential for understanding fine-grained visual descriptions. This raises a critical question:

How to integrate multi-view RGB images, LiDAR point clouds and natural languages for efficient 3D Referring Expression Segmentation?

Current multi-modal fusion approaches [6] in 3D-RES primarily operate at the pixel level without leveraging the object-level semantic information, treating all spatial regions equally rather than adopting collaborative fusion strategies that distinguish between different semantic entities. This leads to a misalignment between multi-modality across images, point clouds and natural language. It is because language inherently describes objects' hierarchical semantics.

To address these limitations, we propose IS-RES, a unified multi-modal fusion framework that explicitly captures pixel-level and instance-level semantics and performs hierarchical cross-modal fusion. Our framework addresses the semantic entanglement problem through two key innovations that work synergistically to enhance language-grounded 3D understanding. First, we introduce **Hierarchical Visual Semantic Decomposition**, a systematic approach that restructures pixel-level processing by establishing semantic hierarchies that distinguish between different granularities of visual information. Our approach leverages Segment Anything (SAM) [31, 59] to automatically extract object instance masks from multi-view images [12], enabling feature extraction at two complementary levels: dense pixel-level CLIP features that preserve fine-grained spatial details and instance-aware mask features that capture semantically coherent object representations while maintaining contextual boundaries. This dual-granularity decomposition directly addresses the feature dilution problem by ensuring that object-specific semantic information remains structurally distinct from background interference throughout the extraction process, providing complementary visual representations that serve as the foundation for subsequent multi-level cross-modal integration. Second, we develop a **Progressive Multi-level Fusion strategy** that transforms cross-modal feature integration by implementing a three-stage hierarchical approach that dynamically balances different feature granularities based on semantic context. This strategy addresses the limitation that current approaches apply uniform weighting across all spatial regions and modalities, failing to accommodate the varying semantic dependencies required by different types of referring expressions. Through Intra-modal Collaborative Integration, we establish complementary interactions between instance-level and pixel-level features within the 2D visual domain using

multi-head attention, creating enriched shared representations. Subsequently, Cross-modal Dynamic Integration employs gating mechanisms to learn context-specific weighting between 2D and 3D features at each superpoint, replacing static summation with adaptive modal balance. Finally, **Language-guided Instance Refinement** leverages linguistic features to refine spatially-selected superpoint candidates and performs attention-based interaction between 2D instance features and unified representations, achieving precise semantic alignment. Together, these innovations enable IS-RES to achieve robust cross-modal alignment that preserves semantic coherence while accommodating the diverse requirements of referring expressions in complex 3D environments. In summary, our main contributions are as follows:

- We propose Hierarchical Visual Semantic Decomposition, a systematic approach that extracts complementary visual representations at multiple granularities, fundamentally addressing the semantic entanglement problem in traditional pixel-level processing.
- We introduce Progressive Multi-level Fusion Strategy, a hierarchical integration framework that achieves adaptive cross-modal feature alignment through intra-modal collaboration, dynamic modal weighting, and language-guided instance refinement.
- We demonstrate that IS-RES achieves state-of-the-art performance on ScanRefer [5] and Multi3DRefer [93] datasets.

This thesis is well-structured and provides an in-depth exploration of our research on 3D Referring Expression Segmentation with a focus on the unified multi-modal approach.

The rest of this thesis is organised as follows:

Chapter 2 provides a comprehensive literature review. It first examines the evolution from 2D Referring Expression Comprehension and Segmentation to their 3D counterparts, analyzing the progression from two-stage to one-stage architectures and the emergence of generalized settings. We then review multi-modal 3D fusion techniques across various perception tasks, with particular focus on instance-level semantic integration strategies. Additionally, we discuss vision foundation models including CLIP and SAM that are critical to our instance-aware feature extraction approach.

Chapter 3 presents our proposed IS-RES framework in detail. We begin by formalizing the 3D-GRES task definition and challenges. We begin by formalizing the 3D-GRES task definition and its challenges in generalized settings. We then introduce our Hierarchical Visual Semantic Decomposition, which employs SAM-guided instance mask extraction to generate dual-granularity features: dense pixel-level CLIP embeddings preserving spatial details and instance-aware features capturing coherent object representations. This approach addresses semantic entanglement through confidence-weighted pixel-to-point association that maintains semantic boundaries during 2D-to-3D projection. Finally, we present the Progressive Multi-level Fusion Strategy, comprising three stages: Intra-modal Collaborative Integration for enriching 2D feature interactions, Cross-modal Dynamic Integration with learnable gating for adaptive 2D-3D feature weighting, and Language-guided Instance Refinement for precise semantic alignment between visual features and linguistic descriptions.

Chapter 4 focuses on the experimental aspects of our research. It outlines the experimental setup, the datasets used, the evaluation metrics, and the comparative methods. We conducted experiments on public datasets and compare with the latest state-of-the-art methods, followed by our ablation studies.

Chapter 5 draws the conclusions based on our research findings, summarise my current research work, highlight the key contributions and discuss the limitations to share possible future research works.

CHAPTER 2

Literature Review

3D Referring Expression Segmentation (3D-RES) extends 3D visual grounding (3D-REC) from coarse-grained bounding box prediction to fine-grained mask generation, demanding pixel-accurate boundary delineation that maintains semantic and geometric coherence for complex language-guided scene understanding. This chapter traces the development from 2D to 3D referring expression tasks, examining the evolution from computationally intensive two-stage methods to efficient end-to-end architectures and the emergence of generalized settings that handle multi-target and zero-target scenarios. To comprehensively understand these architectural advances and their limitations, we systematically review multi-modal 3D fusion techniques that address the semantic deficiencies of point-cloud-only approaches, with particular attention to how vision foundation models—CLIP [58] for semantic understanding and SAM [31, 59] for class-agnostic segmentation—provide powerful priors for instance-aware feature extraction.

However, despite significant advances, current 3D-RES methods predominantly operate on point cloud data alone, overlooking the rich semantic information available in multi-view RGB images. Based on this structured analysis, we identify three critical gaps in existing literature: the predominant reliance on geometric features without sophisticated multi-modal integration, the lack of instance-level semantic awareness in fusion strategies that operate primarily at scene level, and the absence of hierarchical feature decomposition that preserves instance boundaries. These limitations therefore directly motivate the progressive multi-level fusion framework presented in this thesis, which leverages hierarchical visual semantic decomposition and adaptive cross-modal alignment to achieve robust instance-aware segmentation.

2.1 2D Referring Image Segmentation

Building on the task definitions outlined above, we first examine 2D Referring Expression Segmentation (2D-RES) as it provides essential architectural insights for understanding 3D approaches. 2D-RES [41, 61, 80, 13, 82, 69] generates pixel-precise segmentation masks for objects described in natural language, requiring models to align visual features with linguistic semantics at fine granularity. Unlike conventional segmentation tasks that rely solely on visual input, 2D-RES demands sophisticated cross-modal reasoning to interpret spatial relationships, object attributes, and contextual cues embedded in referring expressions. The field has witnessed a clear architectural evolution from two-stage methods that decompose the problem into detection and selection phases to one-stage approaches that directly predict masks through end-to-end optimization. This progression, along with the recent integration of vision-language foundation models, offers valuable lessons for addressing similar challenges in 3d domains.

2.1.1 Two-Stage Methods

Two-stage methods [42, 87, 30], also known as a bottom-up or proposal-based approach, decompose referring segmentation into sequential sub-tasks, first generating comprehensive object proposals through visual analysis, then incorporating linguistic information to identify the target object. This paradigm emerged from the success of region-based object detection [37, 81] frameworks and visual grounding task (also known as comprehension task) [23, 24, 43, 48, 88], adapting their proposal-and-verification structure to incorporate linguistic reasoning. The architecture benefits from leveraging robust pre-trained detectors [19] that have been optimized on large-scale visual datasets, providing strong initial representations for object localization.

To summarise, these two-stage methods decompose the research problem into two distinct but sequential steps.

(1) *Stage 1: Proposal Generation*: The first stage typically employs the off-the-shelf, pre-trained object detector or instance segmentation models to analyze the image and generate a set of candidate instance proposals. These proposals are generated in a language-agnostic manner, extracting all potential candidates objects without concerning any referring expression.

(2) *Stage 2: Grounding and Selection*: The second stage takes the candidate proposals and the language expression encoding as input. It then computes the matching scores between each to select the best proposal candidate among all that corresponds to this language expression as its final output.

Despite their success in related grounding and detection tasks, two-stage methods face significant limitations in RES, as evidenced by the predominant adoption of end-to-end one-stage approaches in recent literature [41]. The primary limitation stems from computational inefficiency and proposal bottlenecks, where separate detector training increases resource requirements, while failed proposal generation creates unrecoverable errors regardless of sophisticated linguistic reasoning. More fundamentally, the discrete proposal paradigm misaligns with segmentation requirements, as referring expressions frequently target object parts, contextual regions, or multi-object groups that cannot be effectively decomposed into independent bounding box proposals. This architectural mismatch explains the divergent evolution: while detection and grounding benefit from coarse spatial reasoning where approximate localization suffices, segmentation demands pixel-level precision that requires integrated visual-linguistic processing throughout the entire pipeline.

2.1.2 One-Stage Methods

One-stage methods [94, 25, 83] directly generate the segmentation masks in a single pass, removing intermediate proposal stages through end-to-end transformer-based architectures [68, 41, 61, 80, 13, 82, 69] aligning vision–language features directly through Transformer decoders that integrate linguistic cues into mask prediction. The paradigm shift was catalyzed by DETR [4], which introduced joint object reasoning through parallel query alignment, dispensing with region proposals and providing a clean architectural template for subsequent

referring segmentation models [51, 35]. Building on this foundation, Vision-Language Transformer (VLT) [13, 82] reformulated referring segmentation as a query-to-pixel attention problem, where language-conditioned object queries enable direct cross-modal alignment between textual descriptions and visual regions. Recent work increasingly leverages vision-language foundation models: CRIS [69] transfers CLIP's multimodal knowledge [58] to pixels via a vision-language decoder and text-to-pixel contrastive learning, while Prompt-RIS [62] bridges CLIP and SAM through bidirectional prompt learning; Universal or reasoning segmentation [80, 10, 33] further expands this direction. While these advances primarily target 2D scenarios, the rich semantic knowledge embedded in these foundation models [58, 31, 59] can be effectively transferred to 3D referring tasks through multi-view image correspondences, motivating a detailed examination of key 2D vision models and their potential integration strategies.

2.1.3 Generalized Referring Segmentation

While traditional 2D-RES methods operate under the restrictive assumption that each referring expression corresponds to exactly one target object, real-world scenarios frequently involve expressions that refer to multiple objects, object parts, or even no objects at all. To address this limitation, Liu et al. [41] introduced Generalized Referring Expression Segmentation (GRES), which extends the task formulation to handle expressions indicating an arbitrary number of target objects, better reflecting real-world scenarios where human expressions naturally refer to zero, single, or multiple objects.

As illustrated in Figure 2.1, GRES addresses three distinct scenarios that traditional RES cannot handle: (1) *multi-target expressions* such as "all people" or "standing people" that refer to multiple objects simultaneously, (2) *zero-target expressions* like "the kid in blue" when no such object exists in the scene, and (3) conventional single-target cases. This expanded formulation requires models to determine target cardinality alongside segmentation accuracy, introducing challenges in both annotation complexity and architectural design. Technically, GRES employs a region-based approach (ReLA) that explicitly models region-region and region-language dependencies through cross-attention mechanisms, where images are

adaptively divided into regions with sub-instance clues rather than simple hard-split partitions, enabling effective multi-object relationship modeling essential for complex referring expressions.

This *generalized* task formulation has influenced subsequent work across both 2D and 3D domains, with Multi3DRefer [93] extending these principles to 3D scenes by providing annotations for multi-object and zero-object referring expressions in point cloud environments, serving as the primary evaluation benchmark for our 3D referring segmentation research.

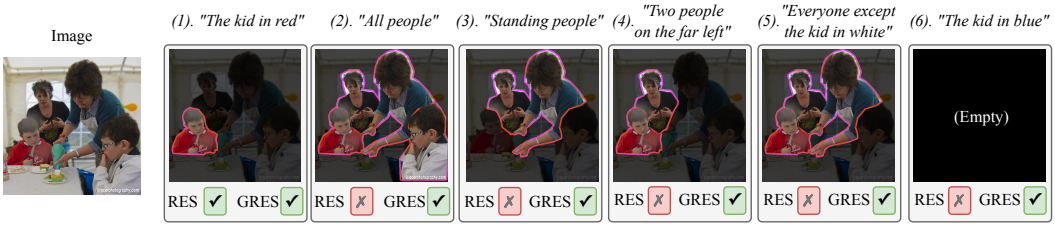

FIGURE 2.1. Comparison between traditional RES and Generalized RES (GRES). While traditional RES only supports single-target expressions, GRES handles multi-target expressions like “all people” and zero-target cases like “the kid in blue” when no matching object exists. Figure adapted from [41].

2.2 2D Vision Model

In reviewing 2D Referring Expression Segmentation tasks, recent works such as CRIS [69], Prompt-driven [62] and other approaches [79, 66, 80, 33, 10] all use large-scale pre-trained models like CLIP [58], SAM [31, 59], and ViT [14]. Image referring segmentation differs fundamentally from instance or semantic segmentation in that it must resolve both visual cues and the semantics embedded in natural-language expressions. Consequently, large-scale vision-language foundation models - most notably CLIP [58], which is trained on hundreds of millions of image-text pairs - are naturally well suited to this task and have already demonstrated strong zero-shot and transfer performance across a wide range of downstream problems. Complementing CLIP on the vision side, the Segment Anything Model [31, 59] provides high-quality, category-agnostic masks and has rapidly become a backbone component for segmentation pipelines.

While the advantages of transferring knowledge from such 2D foundation models are clear for image-based referring segmentation, they are equally valuable for 3D scenes. Indoor 3D datasets [12, 84] are tightly coupled with 2D RGB and depth images: each 3D point can be projected onto one or more image frames, and conversely, 2D observations can be re-projected to reconstruct 3D geometric surfaces. This intrinsic correspondence suggests that CLIP [58], SAM [31, 59], and related models can act as powerful priors for enriching 2D features that are subsequently lifted into 3D space. Accordingly, the remainder of this section reviews these 2D foundation models in detail, explores the possible way to utlise them and then motivates their integration into the proposed 3D referring expression segmentation framework.

2.2.1 CLIP

Contrastive Language–Image Pre-training (CLIP) [58] represents a landmark achievement in multi-modal learning and is one of the most influential pre-trained models across vision–language research. Beyond the strong zero-shot performance of the released checkpoints where researchers tends to use them as "frozen encoders", CLIP's contrastive training objective has also shaped subsequent work on multi-modality. By learning from natural-language supervision at an unprecedented scale, CLIP acquires a rich semantic understanding that supports remarkable zero-shot generalization. As illustrated in Figure 2.2, CLIP adopts a

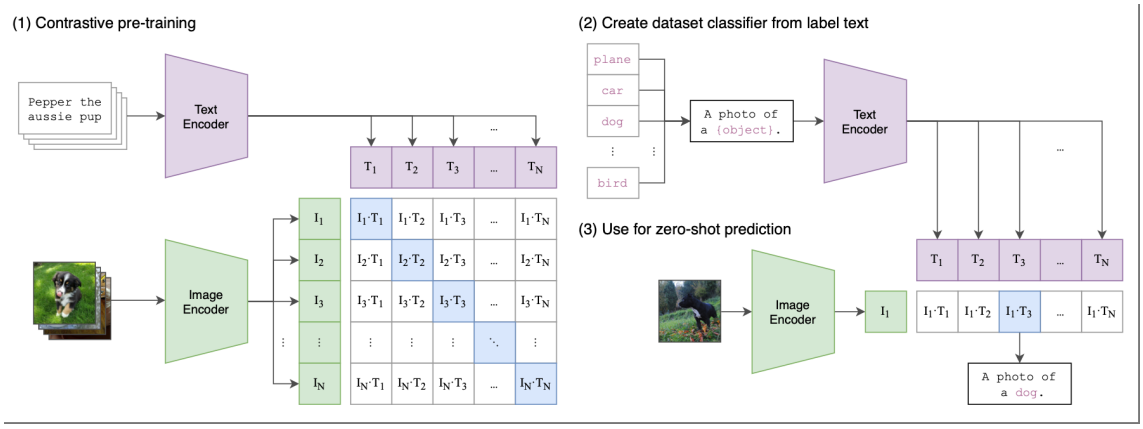

FIGURE 2.2. CLIP Framework [58]

simple yet powerful dual-encoder architecture with two separate pathways that encode visual

and textual inputs before aligning them in a shared feature space. Crucially, the training data are naturally paired images and captions, so the alignment signal is inherent.

- **Image encoder**: The image encoder transforms raw images into fixed-length vector representations. The original CLIP paper explores both ResNet and Vision Transformer (ViT) backbones; ViT [14] variants generally outperform their ResNet counterparts. The default setting pools the tokens in its final layer to yield a single global embedding, though many researchers extract intermediate patch-level features before pooling for downstream use or extend patch-level features to pixel-level representations.
- **Text encoder**: The text branch is a standard multi-layer transformer architecture [68]. It converts a sequence of tokenized text into a fixed-length embedding that captures the sentence's semantics.

The training process of CLIP could be briefly described as a way to teach the model to learn which images and texts belong together and which image-text pairs are irrelevant. For a batch of N image–text pairs (see Figure 2.2), the procedure is:

(1) Each image is passed through the image encoder to obtain embeddings $I_1, \ldots, I_N$; the corresponding texts yield embeddings $T_1, \ldots, T_N$.
(2) The linear projection layers map both modalities into a common d-dimensional space.
(3) The model computes the cosine similarities between every possible image–text pair, producing an $N \times N$ similarity matrix. The diagonal entries in the figure correspond to the N positive pairs that are naturally related to each other, and the off-diagonal entries to the $N^2 - N$ negative pairs.
(4) A contrastive loss simultaneously maximizes similarity for positives pairs and minimizes it for negatives. The model is trained to "pull" matched image–text embeddings together while pushing mismatched pairs apart. Training on hundreds of millions of pairs ensure a feature space that is both semantically aligned and broadly generalizable.

2.2.2 Segment Anything

While CLIP provides a strong foundation for semantic understanding with its powerful image-text representations learned via global contrastive objective, it lacks intrinsic pixel- or instance-level discrimination. In contrast, the Segment Anything Model (SAM) [31, 59] from Meta AI offers promptable, generalized segmentation capabilities and has significantly influenced recent segmentation approaches.

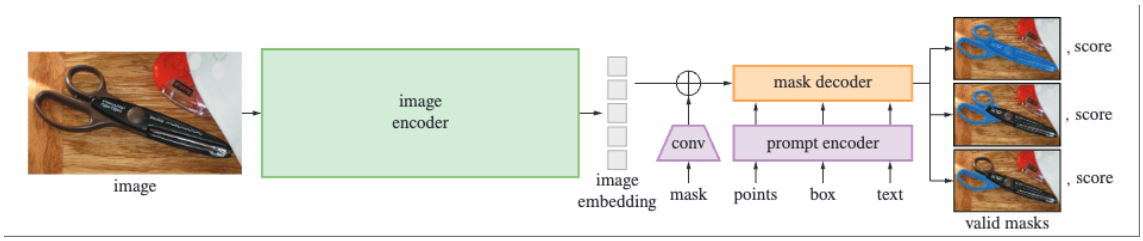

FIGURE 2.3. SAM Framework [31]

As illustrated in Figure 2.3, SAM's architecture was designed for "promptable segmentation", enabling effective, easy, and interactive inference. Its architecture consists of three main components:

(1) **Image Encoder**: A large Vision Transformer (ViT) pre-trained with a masked-autoencoder loss. It takes a high-resolution image as input and encodes it to generate a detailed image embedding that is used throughout interactive prompts.

(2) **Prompt Encoder**: A lightweight network that converts user inputs (sparse prompts like points and boxes or dense prompts like masks) into vectors using positional encoding.

(3) **Mask Decoder**: A lightweight Transformer-based decoder that fuses the frozen image embedding with the prompt embeddings and predicts up to three plausible masks in $\sim$50 ms, enabling interactive exploration.

Different from CLIP [58] which is trained on image-text pairs, SAM is trained without textual information but learns purely from geometric consistency between prompts and masks. SAM's goal is to segment any object indicated by the user's prompt, regardless of category. This

category-agnostic capability makes SAM particularly valuable for generating high-quality instance masks that can serve as spatial priors for downstream applications.

The success of foundation models like CLIP [58] and SAM [31, 59] in 2D referring segmentation naturally motivates their application to 3D scenarios, where the challenges become significantly more complex due to the sparse, irregular nature of point cloud data and the need for sophisticated spatial reasoning. While 2D referring tasks benefit from dense pixel grids, 3D referring expression tasks must address additional complexities including viewpoint variations, scale disparities, and the inherent sparsity of 3D representations. However, the rich semantic knowledge embedded in these 2D foundation models, combined with the natural correspondence between 3D scenes and their multi-view 2D projections, provides a compelling pathway for enhancing 3D understanding through cross-modal knowledge transfer. This potential has driven the development of 3D referring expression comprehension and segmentation methods, which we explore in the following sections, alongside the sophisticated multi-modal fusion strategies required to effectively integrate 2D semantic priors with 3D geometric features.

2.3 3D Referring Expression Comprehension

3D Referring Expression Comprehension (3D-REC), also known as 3D visual grounding [5, 93, 1, 3, 8, 47, 75, 67, 76, 64, 29, 91, 28, 18], aims to localize objects within 3D point cloud scenes using textual descriptions as queries, outputting 3D bounding boxes for referred objects. Both 2D and 3D referring expression comprehension tasks predominantly employ two-stage approaches [5, 93, 1, 3, 8, 67, 76, 64, 91, 28, 18] that separate object detection and language alignment, with fewer works exploring end-to-end architectures [75, 29, 47] for joint visual-linguistic optimization. This architectural distribution reflects the complexity of cross-modal alignment in localization tasks, where explicit object proposal generation often provides more stable performance than direct end-to-end prediction, contrasting with 3D referring expression segmentation tasks (Section 2.4) that favor one-stage approaches.

Representative approaches in 3D-REC demonstrate diverse strategies for cross-modal alignment. BUTD-DETR [29] integrates language-guided attention with bottom-up object proposals, decoding objects directly from visual input while leveraging pre-trained detection outputs as contextual guidance. In contrast, object-centric methods focus on reasoning over individual object point clouds rather than processing entire scenes holistically. NS3D [22] introduces a neuro-symbolic approach that operates on a pre-given set of object instances, parsing natural language into symbolic programs using large language models like Codex, then executing these programs on object-centric features to resolve complex spatial relationships. LARC [15] extends this paradigm by using VoteNet [55] for object detection while incorporating denoising strategies to handle detection uncertainties, emphasizing minimal supervision through language-based constraints as regularization.

Multi-modal integration strategies have proven particularly effective in advancing 3D-REC performance. EDA [75] extends BUTD-DETR [29] by demonstrating successful text-decoupling approaches that extract positional and relational information from referring expressions to guide visual feature alignment, with text decoupling strategies subsequently being adopted in 3D-RES tasks [73, 72, 6, 71]. Recent works have also explored diverse directions: Box2Mask [9] investigates weakly-supervised learning from bounding box annotations, OneFormer3D [32] proposes unified frameworks for multiple 3D understanding tasks, and CIP-WPIS [89] focuses on cross-modal interaction with reduced annotation requirements. Additionally, several approaches [93, 3, 67, 76, 64, 18] leverage multi-view images to enhance spatial understanding through 2D-3D feature correspondence.

The success of multi-modal integration and cross-modal alignment strategies in 3D-REC naturally extends to more challenging tasks requiring pixel-level precision. While 3D-REC establishes object localization through bounding boxes, 3D Referring Expression Segmentation demands precise mask generation that benefits even more substantially from the semantic richness provided by 2D vision-language models. The insights from text-decoupling, neuro-symbolic reasoning, and particularly the effective use of multi-view image correspondences in 3D-REC provide essential foundations for developing sophisticated 2D-to-3D feature integration strategies in segmentation tasks, which we examine in the following sections.

2.4 3D Referring Expression Segmentation

3D Referring Expression Segmentation builds upon the foundation of 2D RES [41, 61, 80, 13, 82, 69, 33, 10] to localize and segment objects in 3D space from natural language descriptions. The task has emerged as an increasingly popular research direction that extends established 3D visual grounding and detection tasks [93, 5, 1, 3, 8, 47, 75] to provide fine-grained segmentation guided by natural language descriptions [27, 42, 73, 72, 6, 90, 56]. Early 3D-RES methods followed a two-stage framework, generating object proposals through pretrained detectors first and then matching them with textual descriptions to compute matching scores for identifying referred targets [27]. This approach was proved to be inefficient and suboptimal [73], prompting researchers to develop end-to-end solutions with different architectural designs. X-RefSeg3D [56] combines linguistic and visual features to create a cross-modal scene graph for interactions based on textual and spatial relations.

3D-STMN [73] introduced superpoints [63, 34] as geometric priors to align with textual features, transitioning from traditional point-level processing to superpoint-level representations that achieve both improved computational efficiency and state-of-the-art results (Figure 2.4). This superpoint-based framework established a new paradigm that has been adopted by subsequent works [72, 26, 6, 71], demonstrating its effectiveness in handling the complexity of 3D scene understanding. Recent approaches such as 3D-STMN and subsequent work [57, 21, 72, 39, 78] adopt one-stage architectures that directly predict segmentation masks through query-based mechanisms. Building upon this superpoint-based architecture, 3D-GRES [72] further extended the task formulation to generalized settings [93, 41], adopting the multi-target and zero-target paradigm established in 2D-GRES (Section 2.1.3) to address scenarios where referring expressions may correspond to multiple objects or no objects. Additionally, several works [57, 21, 39, 78] have explored joint training strategies, simultaneously optimizing 3D-RES alongside related tasks such as visual grounding and object detection to achieve improved overall performance.

However, existing approaches in 3D-RES predominantly rely on geometric features from point clouds which lack sophisticated multi-modal features and fusion strategies that can

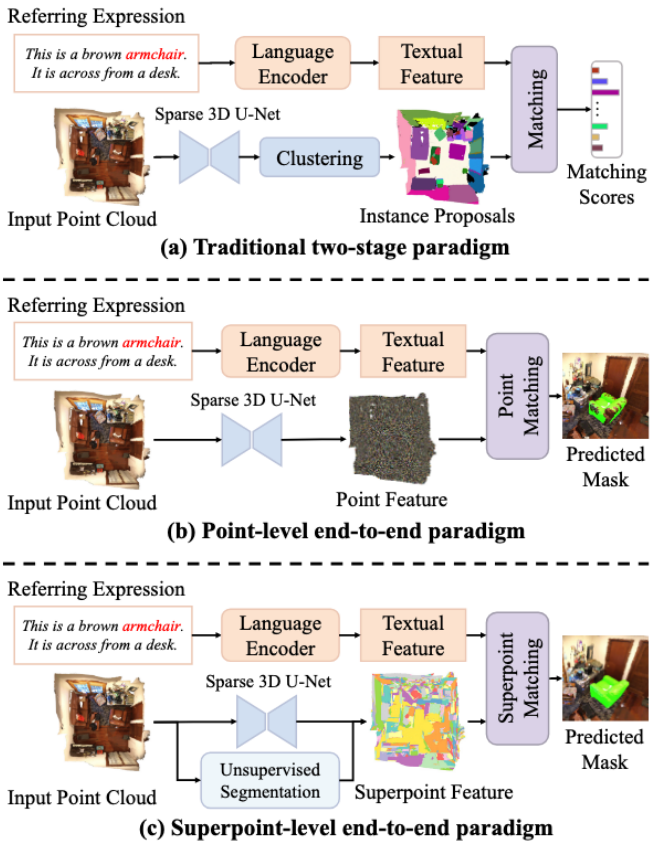

FIGURE 2.4. Evolution of 3D-RES architectures: (a) traditional two-stage paradigm with instance proposals, (b) point-level end-to-end approach, and (c) superpoint-level end-to-end paradigm introduced by 3D-STMN that reduces computational complexity while maintaining performance. Figure adapted from [73].

effectively leverage rich semantic information from pre-trained vision-language models [58, 31]. Therefore, our focus is to explore sophisticated multi-modal fusion strategies by incorporating the merits of both semantic-rich vision-language models [31, 58] and geometry-aware 3D features to achieve more robust and accurate referring expression segmentation in complex 3D environments.

2.5 Multi-modal 3D Fusion

Multi-modal 3D fusion has emerged as a critical technique for enhancing 3D scene understanding by integrating complementary information from different modalities, particularly

combining geometric features from point clouds with semantic features from RGB images, ranging from a variety of tasks including 3D object detection [38, 45, 85, 81, 37], visual grounding [67, 93, 76, 64, 29, 91, 28, 18] , and 3D segmentation [6, 49, 52, 65].

Existing approaches are typically categorized into early fusion, middle fusion, and late fusion strategies based on integration stages [85], where early fusion methods [77, 86] directly enhance input points with image features but suffer from calibration sensitivity, while late fusion approaches [2, 36] integrate modality-specific proposals separately with limited cross-modal interactions during proposal generation. Middle fusion [38, 45] has gained prominence by enabling multi-modal interactions at intermediate stages, with contemporary works aligning point cloud and image features on unified bird's-eye-view representations in detection task [85, 38, 45] and recent efforts exploring multi-view fusion strategies in grounding and segmentation tasks [28, 18, 93, 67, 49, 52] for enhanced robustness. However, most existing fusion strategies operate primarily at the scene level [85] without distinguishing foreground objects from background noise, while recent works have identified the importance of leveraging fine-grained instance-level information and multi-attributes interactions for effective object disambiguation [67, 76].

In contrast to these approaches, our work develops a progressive multi-level fusion framework that leverages SAM-guided dual-granularity features [31, 58] to incorporate precise instance boundary knowledge, enabling dynamic weighting mechanisms and fine-grained instance-attribute interactions that address the inherent limitations of scene-level fusion approaches.

CHAPTER 3

Methodology

3.1 Problem Formulation

We address the task of *Generalised Referring Expression Segmentation (GRES)* in 3D indoor scenes, where the target objects to be segmented can be multiple objects, a single object, or no objects when the referring expression cannot be grounded in the scene [93, 72]. Given a point cloud scene $P \in \mathbb{R}^{N_p \times C}$ with N_p points, C denotes the channel dimension and a natural language utterance $\mathcal{U}$ describing target object(s) and their spatial relations, the goal is to produce a binary segmentation mask $\mathcal{M} \in \{0, 1\}^{N_p}$ that identifies all points belonging to the referred object(s).

3.2 Preliminary

3.2.1 Scene Representation

We conduct our study on 3D indoor scenes from the ScanNet dataset [12], where each scene is represented by an RGB-colored point cloud $\mathcal{P} \in \mathbb{R}^{C \times 6}$, with each point defined by its 3D coordinates (XYZ) and RGB color. A typical scene contains approximately 100,000 points along with 20–50 posed RGBD frames, which provide depth maps, camera intrinsics, and camera poses.

These multi-view RGB images enable the extraction of dense semantic features and instance-aware cues using pretrained 2D vision-language models. Later, these features are projected into the 3D space and associated with the underlying geometry. In addition, each scene is

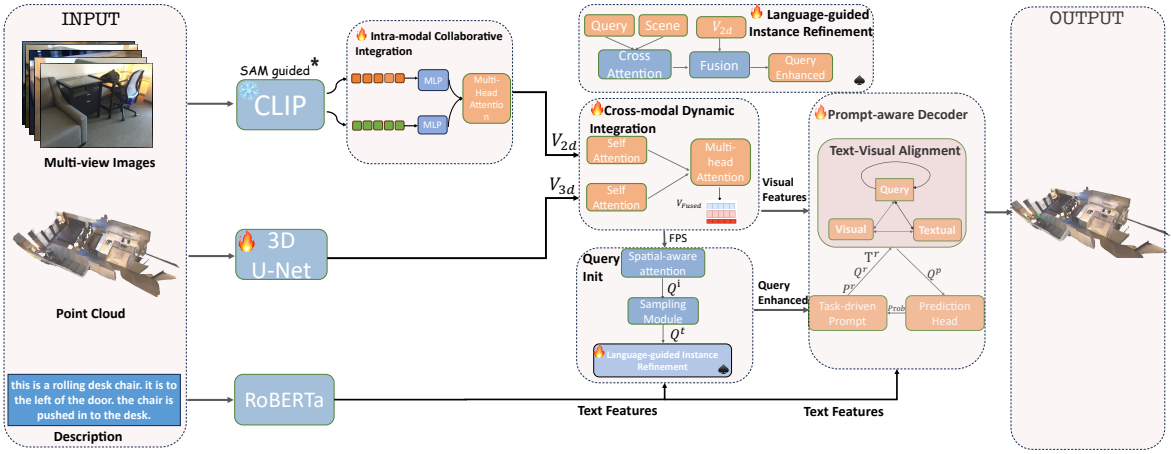

FIGURE 3.1. **Pipeline of our proposed IS-RES framework.** The framework processes three input modalities: point clouds via 3D U-Net, text via RoBERTa, and multi-view images via CLIP guided by SAM* through our Hierarchical Visual Semantic Decomposition (Section 3.3.3 and Figure 3.2). Object queries are initialized via Farthest Point Sampling (FPS) on point cloud. After cross-modal feature integration and language-guided sampling, our Language-guided Instance Refinement enhances selected queries through scene context awareness and 2D semantic fusion. Enhanced queries are decoded via a 6-layer prompt-aware decoder for final 3D referring expression segmentation.

annotated with natural language expressions [5, 1, 93] that refer to specific objects, forming the core input $\mathcal{U}$ in the GRES task. Together, these multi-modal inputs support learning to segment 3D regions based on free-form linguistic descriptions.

3.2.2 Superpoint Representation

To organize the 3D point cloud into semantically meaningful units and reduce computational redundancy, we adopt geometric-based oversegmentation to generate superpoints [17, 63]. Each superpoint [34] is a spatially coherent cluster of points with similar local geometry. This representation is widely used in 3D scene understanding [63, 73, 72, 49] and allows feature extraction, fusion, and prediction to operate at the superpoint level rather than per-point, offering a balance between resolution and efficiency.

3.2.3 Pipeline Overview

Our IS-RES framework builds upon prior 3D GRES pipelines [73, 72, 6] by establishing a systematic encoder-decoder architecture that integrates geometric, visual, and linguistic information through cross-modal alignment mechanisms, while addressing fundamental limitations in semantic representation and modal integration strategies. The encoding stage employs a pre-trained Sparse 3D U-Net [17] to extract point-wise geometric features from voxelized 3D point clouds, while simultaneously utilizing a frozen RoBERTa encoder [44] to generate contextualized language embeddings from natural language descriptions. Following our Hierarchical Visual Semantic Decomposition approach (Section 3.3), the framework incorporates SAM [31] for instance-aware mask generation from multi-view images, facilitating dual-granularity visual representations through both dense pixel-level CLIP features that preserve spatial details and semantically coherent instance-level features that maintain object boundaries. Subsequently, the system performs superpoint pooling and Furthest Point Sampling (FPS) [50] to establish representative spatial anchors, from which language-refined queries [72] are initialized through our Progressive Multi-level Fusion Strategy 3.4 that systematically integrates multi-granularity visual embeddings across 2D-3D modalities via intra-modal collaboration, cross-modal dynamic weighting, and language-guided instance refinement. The decoder architecture inherits from 3D-GRES frameworks [72, 6] with stacked Transformer layers enhanced by task-driven prompt mechanisms adapted from IPDN [6] that address intent ambiguity by dynamically prioritizing text-relevant queries, enabling the model to focus on semantically important regions rather than applying uniform attention across all spatial locations during the segmentation prediction process.

3.3 Hierarchical Visual Semantic Decomposition

3.3.1 Language Encoding

We employ the pre-trained RoBERTa [44] to encode the referring expressions $\mathcal{U}$ into text tokens $\mathcal{T} \in \mathbb{R}^{N_T \times D_T}$, where N_T denotes the number of tokens, and D_T denotes the embedding

dimension. To facilitate cross-modal alignment in the decoder, we employ a linear projection to transform $\mathcal{T}$ into a unified feature space:

$$T = \mathcal{T}W_T, \tag{3.1}$$

where $W_T \in \mathbb{R}^{D_T \times D}$ represents the learnable parameters. To establish precise supervision signals for cross-modal alignment, we adopt the scene graph parser [60, 74] following [75, 72] to decompose referring expressions into structured semantic components. Specifically, each expression is parsed into five distinct categories: `Main object` (primary target entity), `Attributes` (visual properties such as color and shape), `Spatial relations` (geometric relationships between objects), `Pronouns` (referential terms), and `Auxiliary objects` (contextual entities for spatial reasoning). This structural decomposition enables fine-grained supervision of vision-language correspondence. Following the baseline methodology [72], normalized positive maps are constructed for each semantic component to provide token-level supervision signals in the contrastive semantic alignment loss during decoding process.

3.3.2 Dense Multi-view Feature Encoding

Current approaches for 3D-RES predominantly rely on geometric features from point clouds, which lack sufficient semantic richness necessary for robust language alignment [6]. To bridge this semantic gap, we leverage pre-trained CLIP models [58] that provide rich visual-semantic associations learned from large-scale image-text corpora. We process multi-view RGB images $\{I_i\}_{i=1}^{N_I}$ through the CLIP visual encoder to extract patch-level semantic features, which are subsequently upsampled via bilinear interpolation to obtain dense pixel-level representations $\{F_i^{img} \in \mathbb{R}^{H \times W \times C_I}\}_{i=1}^{N_I}$. These 2D semantic features are projected into 3D space using camera parameters [89, 53, 92, 7, 85, 54], where each pixel (u, v) with valid depth $d(u, v)$ is transformed to world coordinates through:

$$\mathbf{p}_{\text{world}} = \mathbf{T}_{\text{cam}}\mathbf{K}^{-1} \begin{bmatrix} u \cdot d(u,v) \\ v \cdot d(u,v) \\ d(u,v) \end{bmatrix}. \tag{3.2}$$

The projected pixel coordinates serve as sphere centers in 3D space, where all point cloud points within each spherical region inherit the corresponding pixel feature, and these point-level features are subsequently aggregated to obtain superpoint-level representations that serve as the shared space for multi-modal fusion. However, since pixels lack explicit instance boundaries and spatial context awareness, the geometric projection and subsequent superpoint aggregation process introduces noise where semantically distinct objects become entangled within shared superpoint representations. This semantic entanglement issue motivates the need for explicit instance-level semantic enhancement that preserves object boundaries during the projection and aggregation process.

3.3.3 Instance-aware Semantic Enhancement

To overcome the semantic entanglement inherent in dense pixel features, we propose an instance-aware semantic enhancement mechanism that explicitly leverages object segmentation to maintain semantic purity throughout the multi-view fusion process. We first employ SAM [31] to automatically generate instance segmentation masks across all multi-view RGB images, producing sets of object masks $\mathcal{M} = \{M_1, M_2, \ldots, M_N\}$ for each frame $I \in \mathbb{R}^{H\times W\times 3}$ without requiring manual annotations or task-specific training. Each generated mask $M_i \in \{0,1\}^{H\times W}$ is accompanied by predicted IoU and stability scores, enabling quality-based filtering to retain only high-confidence instance segmentation that provides reliable object boundaries for subsequent feature extraction. Since binary masks with hard boundaries would cause discontinuous feature weighting and lose information at object edges, we apply Gaussian blur to generate soft masks:

$$\tilde{M_i} = \mathcal{G}_\sigma * M_i \tag{3.3}$$

where $\mathcal{G}_\sigma$ is a 2D Gaussian kernel with standard deviation σ, and $*$ denotes spatial convolution. Rather than cropping individual instances which would discard valuable contextual information, we feed the complete RGB image through the CLIP visual encoder [58] to obtain spatial feature tokens $f_{spatial} \in \mathbb{R}^{N_{patch}\times D}$ that encode local semantic patterns while preserving full spatial context. We resize the soft mask $\tilde{M}i$ to the CLIP spatial token grid resolution via

bilinear interpolation to obtain patch-level mask coverage $m_j \in [0, 1]$ for each token position j. To account for segmentation quality variations, we weight each token by combining its spatial mask coverage with the SAM prediction confidence:

$$w_j = m_j \cdot q_i \tag{3.4}$$

where $q_i = \text{IoU}_i \times \text{Stability}_i$ is the quality score from SAM's predicted IoU and stability score for mask i. This formulation ensures that features from unreliable segmentations contribute less to the final representation. We then compute instance-specific features through mask-weighted pooling:

$$f_{inst} = \frac{\sum_{j=1}^{N_{patch}} w_j \cdot f_j^{spatial}}{\sum_{j=1}^{N_{patch}} w_j} \tag{3.5}$$

where w_j represents the soft weight at token position j, effectively extracting semantic features from the instance region while maintaining smooth transitions at boundaries. These instance-aware features $f_{inst} \in \mathbb{R}^D$ capture semantically coherent object representations while preserving contextual information from the complete image input. When these instance-aware features are projected to 3D space following the same spherical querying and superpoint aggregation methods, each feature corresponds to a semantically coherent object region rather than arbitrary pixel patches, reducing inter-object semantic interference and preserving clear object boundaries in the final superpoint representations. By combining both dense and instance-aware features, our multi-view semantic encoding provides complementary perspectives: dense features capture fine-grained local patterns while instance features preserve object-level semantic coherence, together forming a comprehensive representation for language grounding. These 3D visual features, combined with the complementary 2D instance-aware and dense features, form the foundation for our Progressive Multi-level Fusion Strategy that systematically integrates multi-modal representations for language-grounded 3D understanding.

3.3.4 3D Scene Encoding and Superpoint Pooling

We employ the voxelization operation [11] on the point cloud and utilize a U-Net style backbone [17] to extract point-wise features $\mathbf{F}_p \in \mathbb{R}^{N_p \times C}$, where C denotes the channels. To

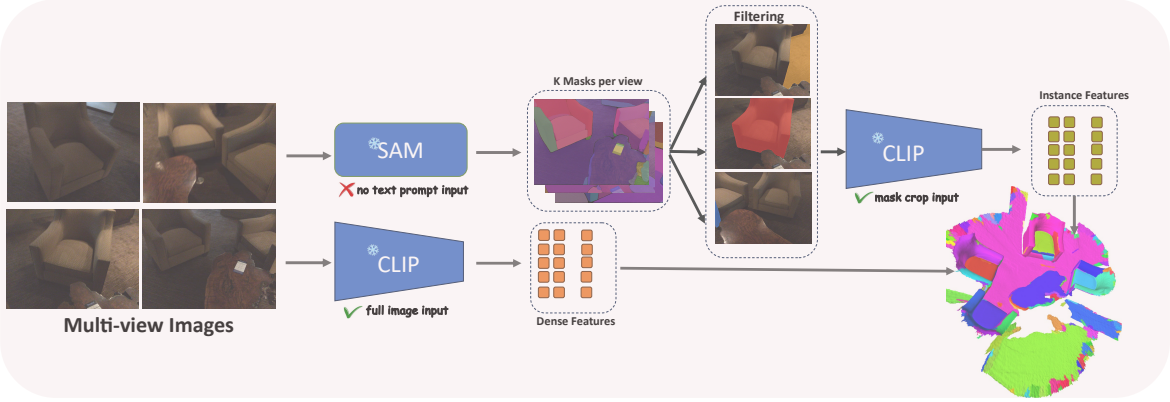

FIGURE 3.2. **Overview of our Hierarchical Visual Semantic Decomposition**: We employ the SAM [31] to segment the instances segmentation masks for each multi-view images without requiring annotations and each mask is then filtered by quality before encoding with CLIP [58] to obtain its instance-level and pixel-level features. These multi-granularity features are then subsequently projected to the point cloud and aggregated into superpoints representations.

reduce computational complexity, we feed these features into a superpoint pooling layer that leverages pre-computed superpoint [34] introduced in 3.2.2. Following [63, 72], we generate N_S superpoints through geometric clustering and aggregate point-wise features within each superpoint via average pooling to obtain $F_{3d} \in \mathbb{R}^{N_S \times C}$. To enable cross-modal alignment, we project the superpoint features into a shared embedding space using an adapter:

$$V_{3d} = \text{Adapter}(F_{3d}), \tag{3.6}$$

where $V_{3d} \in \mathbb{R}^{N_S \times D}$ represents the projected 3D visual features that serve as spatial anchors for subsequent multi-modal fusion. These features, along with textual embeddings, are used to initialize sparse queries for cross-modal alignment processing.

3.4 Progressive Multi-level Fusion

After obtaining rich semantic embeddings from both dense multi-view projections and instance-aware feature extraction 3.3.3, we introduce a multi-level fusion strategy that systematically integrates these complementary 2D representations with 3D geometric features. Our fusion framework operates hierarchically: first aggregating multiple branches of 2D semantic

features into a unified representation, then performing dynamic cross-modal fusion that adaptively balances the contributions from different modalities based on their local relevance. This progressive fusion design ensures that both fine-grained instance semantics and spatial geometric structures are preserved throughout the integration process, ultimately producing a robust multi-modal representation for downstream query-based reasoning.

3.4.1 Intra-modal Collaborative Integration

Building upon the hierarchical visual semantic decomposition established above 3.3, we address the challenge of effectively integrating complementary 2D representations through a collaborative fusion mechanism that leverages multi-head attention to dynamically balance dense spatial features and instance-aware semantics. We process the dense pixel-level features $\mathbf{F}_{\text{dense}}$ and instance-aware features $\mathbf{F}_{\text{inst}}$ as separate branches at the superpoint level, where each branch preserves its distinctive semantic characteristics through dedicated neural pathways before collaborative integration. Rather than naive concatenation that treats all semantic components equally, we decompose these complementary features into independent processing streams that maintain their semantic integrity while enabling dynamic interaction. These processed features are integrated through multi-head attention that learns to dynamically weight different semantic aspects based on local context, enabling the model to emphasize dense spatial details versus instance-level coherence depending on the specific requirements of the referring expression:

$$V_{2d}^{fused} = \text{MultiHeadAttn}(\text{MLP}_{\text{dense}}(\mathbf{F}_{\text{dense}}), \text{MLP}_{\text{inst}}(\mathbf{F}_{\text{inst}})) \tag{3.7}$$

This collaborative integration produces a unified 2D representation that preserves both fine-grained spatial details and instance-level semantic coherence, establishing a robust foundation for subsequent cross-modal dynamic integration with the 3D geometric features V_{3d}.

3.4.2 Cross-modal Dynamic Integration

While element-wise addition provides a baseline strategy for combining the unified 2D representation V_{2d}^{fused} with 3D geometric features V_{3d}[6], this approach fails to account for the varying reliability and contextual relevance of different modalities across spatial locations [67]. To address this limitation, we introduce a spatially-adaptive weighting mechanism that learns to dynamically balance modal contributions by jointly examining both geometric and semantic characteristics at each superpoint location. Specifically, the fusion module processes the concatenated features $[V_{2d}^{fused}; V_{3d}]$ through a lightweight neural network that predicts optimal blending weights for each modality:

$$V_{\text{unified}} = w_{2D} \odot V_{2d}^{fused} + w_{3D} \odot V_{3d} \tag{3.8}$$

This spatially-adaptive weighting enables the model to emphasize 3D geometric features in regions where spatial relationships and geometric constraints are crucial, while prioritizing 2D semantic features in areas rich with visual attributes such as color, texture, and appearance details, where the blending weights w_{2D} and w_{3D} are learned parameters that adapt to local geometric and semantic characteristics, effectively leveraging the complementary strengths of both modalities to produce robust multi-modal representations for subsequent language-guided processing.

3.4.3 Sparse Query Initialization

Following the established sparse query generation framework from prior work [72, 6], we initialize queries from the unified multi-modal representation $V_{\text{unified}} \in \mathbb{R}^{N_S \times D}$ obtained from cross-modal dynamic integration and textual embeddings $T \in \mathbb{R}^{N_T \times D}$. The initialization process employs a hierarchical selection and refinement strategy that transforms dense superpoint representations into sparse, semantically aware queries while maintaining computational efficiency for decoder processing. We first apply farthest point sampling (FPS) [50] on superpoint coordinates to ensure spatially diverse coverage across the scene:

$$Q_{\text{seed}} = V_{\text{unified}}[\text{FPS}(P_{\text{sp}})], \tag{3.9}$$

where $P_{\text{sp}} \in \mathbb{R}^{N_S \times 3}$ denotes the superpoint coordinates and $Q_{\text{seed}} \in \mathbb{R}^{2m \times D}$ represents the geometrically sampled seed features. The seed features are enhanced through spatial-aware attention [6] for local context integration and refined via language-guided selection that identifies semantically relevant candidates based on text-visual alignment scores. The refined seeds are transformed into initial sparse queries through:

$$Q_0 = \text{MLP}(\text{LangSelect}(\text{SpatialAttn}(Q_{\text{seed}}), T)), \tag{3.10}$$

where $Q_0 \in \mathbb{R}^{m \times D}$ ($m \ll N_S$) denotes the sparse initial queries that provide geometrically diverse and linguistically-aligned representations, which act as starting points for subsequent instance-aware refinement processing.

3.4.4 Language-guided Instance Refinement

While the unified multi-modal representation V_{unified} provides a comprehensive foundation for cross-modal reasoning, performing complex instance-aware interactions directly across all superpoint would impose prohibitive computational overhead during training and inference [63, 72]. Building upon the sparse query initialization established above 3.4.3, we employ a progressive refinement strategy that leverages language-guided selection [72, 6] to identify the most semantically relevant queries from Q_0 for detailed instance-aware processing. Specifically, we apply cross-attention mechanisms between the initial queries and textual embeddings T to compute language-visual relevance scores, selecting the top-k most semantically aligned queries $Q_{\text{selected}} \in \mathbb{R}^{k \times D}$ where $k \ll m$ for subsequent instance-guided enhancement. At this computationally efficient scale, we introduce instance-guided enhancement that leverages the 2D instance features $\mathbf{F}_{\text{inst}}$ extracted through our SAM-guided approach 3.3 to enrich the selected queries through cross-attention mechanisms, enabling each query to discover semantic associations with scene-wide instance information while maintaining computational tractability. This progressive refinement approach effectively balances computational efficiency with semantic precision, producing instance-aware queries that benefit from both global scene understanding and fine-grained object-level semantics for subsequent cross-modal reasoning in the decoder framework [6].

3.5 Loss

The loss function of our method consists of five components, following the design principles from [72, 6, 75]. The first component is the instance segmentation loss $\mathcal{L}_{seg}$ applied to queries corresponding to target instances [72]:

$$\mathcal{L}_{seg} = \mathcal{L}_{BCE}(\hat{M}, M_{gt}) + \mathcal{L}_{Dice}(\hat{M}, M_{gt}) \tag{3.11}$$

where $\hat{M}$ denotes the predicted mask, M_{gt} represents the ground truth mask, $\mathcal{L}_{BCE}$ is the binary cross-entropy loss, and $\mathcal{L}_{Dice}$ optimizes geometric overlap.

The second component $\mathcal{L}_{conf}$ supervises the confidence estimation using IoU as the target:

$$\mathcal{L}_{conf} = \text{MSE}(s_{pred}, \text{IoU}(\hat{M}, M_{gt})) \tag{3.12}$$

where s_{pred} is the predicted confidence score, applied only when IoU > 0.5.

The third component is the query indication loss $\mathcal{L}_{ind}$. Following DETR [4], we employ a binary classification loss to determine query validity in the decoder:

$$\mathcal{L}_{ind} = \text{BCE}(\hat{y}_{ind}, y_{ind}) \tag{3.13}$$

where $\hat{y}_{ind}$ represents the predicted indication probability and $y_{ind} \in \{0, 1\}$ indicates whether the query corresponds to a valid target.

The fourth component $\mathcal{L}_{sample}$ encourages semantic-aware point sampling using Focal Loss [40]. Traditional geometric sampling methods like Farthest Point Sampling (FPS) [50] ensure spatial coverage of the superpoint candidates but ignore semantic relevance to language queries. To address this limitation, we employ a sampling quality loss that trains the model to prioritize semantically relevant points:

$$\mathcal{L}_{sample} = \text{FocalLoss}(\hat{s}_{sample}, s_{target}) \tag{3.14}$$

The target scores are computed based on the spatial proximity between sampling points and object centroids, encouraging the selection of points near target instances.

The fifth component $\mathcal{L}_{align}$ ensures proper vision-language alignment through this bidirectional contrastive learning loss adopted from [75, 72].

The final loss $\mathcal{L}$ is calculated as the sum of the losses mentioned above:

$$\mathcal{L}_{total} = \lambda_{seg}\mathcal{L}_{seg} + \lambda_{conf}\mathcal{L}_{conf} + \lambda_{ind}\mathcal{L}_{ind} + \lambda_{sample}\mathcal{L}_{sample} + \lambda_{align}\mathcal{L}_{align} \tag{3.15}$$

where $\lambda_{seg}, \lambda_{conf}, \lambda_{ind}, \lambda_{sample}, \lambda_{align}$ are hyperparameters.

CHAPTER 4

Experiments

4.1 Experimental Setup

In this section, we will elaborate on the experimental setup, including the datasets used, experimental evaluation metrics and the comparison methods. Additionally, we will also present our evaluation results and conduct ablation studies to verify the effectiveness of each component we proposed.

4.1.1 Datasets and Evaluation Metrics

We conduct experiments on two datasets that provide text annotations for ScanNet [12] 3D scenes.

4.1.1.1 ScanRefer

We utilize the ScanRefer dataset [5] to evaluate our method for 3D referring expression segmentation (3D-RES). ScanRefer is the first large-scale dataset for 3D object localization in RGB-D scans using natural language descriptions. The dataset consists of 51,583 natural language expressions encompassing 11,046 objects across 800 ScanNet scenes [12]. Each expression provides a free-form natural language description of one target object within a 3D scene (point cloud), enabling visual grounding of language to 3D geometry.

The dataset builds upon Scannet [12], which provides RGB-D scans of indoor scenes with semantic labels in each 3D scene. Our implementation uses both the original ScanNet 3D point

cloud data and 2D visual features encoding from our proposed decomposition approach 3.3to improve multi-modal understanding.

The evaluation can be broken down into two subsets. The unique subset contains samples where only one object of this category matches the description. The multiple subset contains ambiguous cases with multiple objects of the same category requiring fine-grained distinction (with distractors).

The evaluation metrics of ScanRefer include:

(1) Mean Intersection over Union (mIoU): Measures the average overlap between predicted and ground truth segmentation masks.
(2) Acc@0.25: Accuracy at IoU threshold of 0.25, indicating successful localization with moderate precision.
(3) Acc@0.5: Accuracy at IoU threshold of 0.5, indicating successful localization with high precision

4.1.1.2 Multi3DRefer

Multi3DRefer [93] dataset is used to evaluate our model's performance on 3D Generalized Referring Expression Segmentation (3D-GRES) task. This dataset extends beyond traditional 3D-RES by handling cases where the number of the target objects referenced by the natural language can be arbitrary (zero,one, or multiple objects).

The dataset consists of 61,926 language descriptions with the following:

(1) 51,583 descriptions directly inherited from ScanRefer for consistency
(2) 6,688 descriptions that match zero targets (requires the model to correctly identify no-target scenarios)
(3) 13,178 descriptions that match multiple targets (requires the model to handle multi-object segmentation)
(4) Remaining descriptions that match single targets (traditional referring expression scenarios)

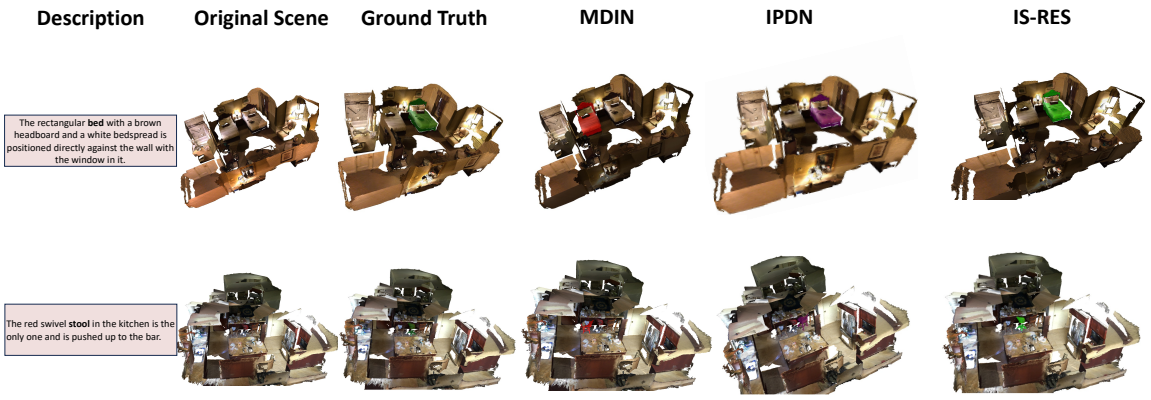

FIGURE 4.1. Qualitative Result to compare our method IS-RES with the state-of-art MDIN [72] and IPDN [6]. Overall, our method produces more accurate segmentation. While IPDN achieves comparable results in general, it often shows noisy predictions near object boundaries.

The evaluation metrics remain consistent with ScanRefer (mIoU, Acc@0.25, Acc@0.5) but include special handling for zero-target cases. When the text refers to zero targets, the sample's mIoU is set to 1.0 if the model correctly identifies this no-target scenario, otherwise 0.0:

(1) zt_w_d/zt_wo_d: Zero-target scenarios with/without distractors.
(2) st_w_d/st_wo_d: Single-target scenarios with/without distractors.
(3) mt: Multi-target scenarios

4.1.2 Training Details

4.1.2.1 Details of running SAM

When using SAM [31] segmentation on all 2D frames from ScanNet [12] dataset, we employ the Vit-H SAM [31], which is the default public model of SAM. Next, we resize each RGB image frame to a resolution of 240×320 to match depth image dimensions. For automatic-SAM segmentation, we set the prediction IoU threshold $\theta_{\text{IoU}} = 0.85$ and stability score

threshold $\theta_{\text{stability}} = 0.90$ to ensure high-quality segmentation masks. We use 64 points per side for sampling density and allow a maximum of 100 masks per frame. To filter low-quality masks, we apply area ratio filtering to retain masks with area ratios between 0.02% and 80% of the total image area, followed by quality score ranking based on $q = \text{IoU}_{\text{pred}} \times \text{stability}_{\text{score}}$ to select the top-30 highest quality instance masks per frame. These filtered masks $\mathcal{M} = \{m_1, m_2, \ldots, m_K\}$ are then used for soft mask-based CLIP feature encoding to extract instance-aware visual representations.

4.1.2.2 Details of encoding with CLIP

After obtaining filtered instance masks $\mathcal{M}$, we adopt an efficient soft mask-based strategy using the ViT-L-14-336 CLIP model [58]. Instead of cropping individual instances, we employ an "encode-once, mask-average-multiple" approach where each entire RGB image is encoded once to obtain spatial patch tokens, motivated by both computational efficiency and empirical analysis showing that individual instance cropping yields highly similar feature distributions with cosine similarities exceeding 0.90 across selected scene samples. For each SAM mask, we compute soft weight matrices by applying Gaussian blur [16] to create smooth mask boundaries, then perform mask-weighted averaging pooling to generate 1024-dimensional instance features as described as 3.3.3. The extracted instance features are subsequently projected to 3D superpoints through ball query with radius of 0.04m and 25 nearest neighbors.

4.1.2.3 Implementation details

We use the pre-trained Sparse 3D U-Net [17] to extract point-wise features from 3D scenes and the pre-trained RoBERTa [44] to extract features from text. For the 2D semantic features, we employed the frozen SAM [31] to segment the instance masks for images from ScanNet [12] and the CLIP [58] to extract dense semantics from 2D images [6] and the instance-aware semantics after SAM's mask. These two pre-trained vision models are frozen entirely in our framework.

For the rest of the networks, we kept similar settings with [72, 6] applies the PolyRL strategy to adjust the learning rate starting from 0.0001, with a decay power of 4.0. The batch size is

TABLE 4.1. Comparison of the 3D-GRES methods on Multi3DRefer. Acc@0.25 and Acc@0.5 refer to the accuracy under IoU thresholds 0.25 and 0.5 respectively. ZT, ST, and MT represent zero target, single target, and multiple targets, respectively. The left and right sides of the "/" represent the situations with and without distractor objects, respectively. The results for MDIN† [72] and IPDN† [6] are obtained from our own reproductions. For others, we used the mIoU and accuracy metrics based on the values provided in their respective papers.

Method	Acc@0.25				Acc@0.5				mIoU
	ZT	ST	MT	All	ZT	ST	MT	All	
ReLA [41]	36.2 / 72.7	48.3 / 83.4	73.0	61.8	36.2 / 72.7	20.4 / 65.5	42.4	37.4	42.8
M3DRef-CLIP [93]	39.2 / 81.6	50.8 / 77.5	66.8	55.7	39.2 / 81.6	29.4 / 67.4	41.0	37.5	37.4
3D-STMN [73]	42.6 / 76.2	49.0 / 77.8	68.8	60.4	42.6 / 76.2	24.6 / 69.2	43.9	40.9	43.0
MDIN [72]	47.9 / 78.8	55.5 / 84.4	76.3	67.0	47.9 / 78.8	29.5 / 71.7	46.8	44.7	47.5
MDIN† [72]	22.2 / 67.2	54.2 / 86.6	76.5	65.4	22.2 / 67.2	27.0 / 71.6	45.6	41.8	45.8
IPDN [6]	39.4 / 84.1	61.5 / 88.9	**79.6**	71.5	39.4 / 84.1	34.7 / 79.5	52.1	50.0	51.7
IPDN† [6]	36.8 / 81.1	59.9 / 86.8	77.9	69.7	36.8 / 81.1	35.3 / 78.9	51.1	49.7	50.8
Is-RES	**47.9 / 86.0**	**62.9 / 88.4**	78.9	**72.3**	**47.9 / 86.0**	**39.5 / 82.2**	**52.9**	**53.4**	**53.5**

set to 16. The number of queries m is set to 128, and the decoder [73] consists of 6 layers with $D = 256$. We use the Farthest Point Sampling (FPS) [50] to first select 256 seed superpoints, which are then filtered with text correlation scores to 128 queries. The maximum number of language tokens is set to 8 with a maximum description length of 78 characters. We employ the AdamW optimizer [46] with weight decay of 5×10^{-4}. The loss function combines multiple components with weights: $\lambda_{seg}, \lambda_{conf}, \lambda_{ind}, \lambda_{sample}, \lambda_{align} = 2.0, 0.5, 0.1, 5.0, 0.1$ reflect the relative importance of each supervision signal with details specified at sec 3.5. All experiments are conducted using the PyTorch framework on two NVIDIA GeForce RTX 4090 GPU.

4.2 Comparison to Baseline

To demonstrate the effectiveness of our model in the 3D express referring segmentation task, we compare Is-RES against state-of-the-art methods on the Multi3DRefer [93] and ScanRefer[5] validation set, as shown in Table 4.1 and Table 4.2. We first discuss the

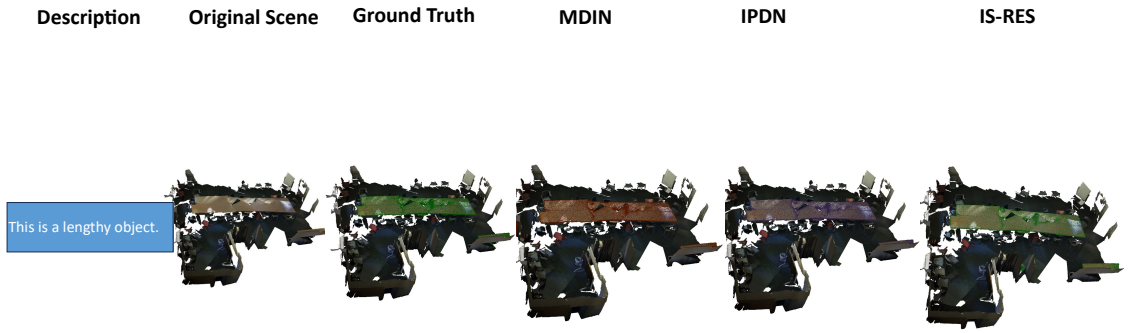

FIGURE 4.2. More qualitative Result to compare our method IS-RES with the state-of-art MDIN [72] and IPDN [6].

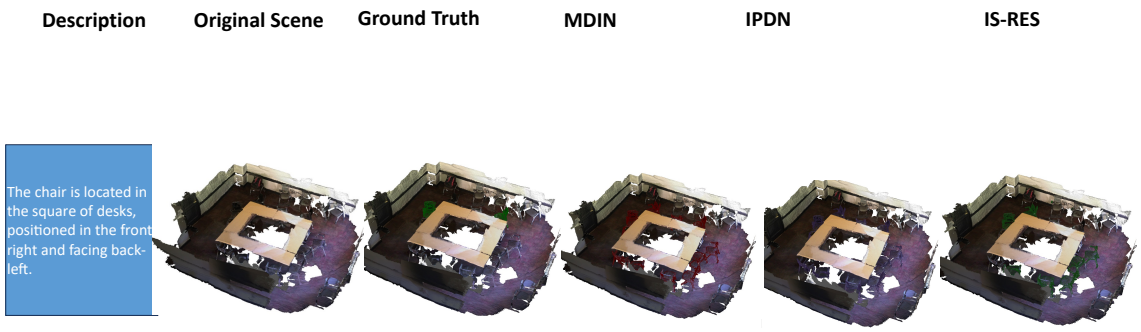

FIGURE 4.3. Failure case to show limitations of current approaches including our IS-RES method. These challenging scenarios reveal common failure modes across all methods, including difficulties with severely occluded objects, ambiguous spatial relationships, and complex multi-object configurations.

results on 3D-GRES, which is the harder benchmark with "multi-target" and "zero-target" incorporated.

To evaluate the effectiveness of our proposed method for the *3D Generalised Referring Expression Segmentation* setting, we compare against both classical and recent state-of-the-art baselines on this 3D-GRES benchmark [93], in particularly on two representative methods:

- MDIN [72]: The first work that designed specifically for 3D-GRES, a 3D-only method, built on 3D-STMN[73]. It does not use any 2D priors and therefore serves as our primary baseline.
- IPDN [6]: A recent method published during our project development. IPDN extends MDIN [72] by injecting multi-view 2D features. Due to its structural similarity and

publicly available codebase, we use it as an extended baseline and reference for implementation.

Our Is-RES achieves the highest overall performance on Multi3DRefer [93] with 53.5 mIoU, surpassing the previous best method IPDN [6] by 2.7 points and substantially outperforming MDIN [72] by 7.7 points. The large performance gap between ours and MDIN [72] highlights the limitations of 3D-only approaches in understanding complex visual attributes that are essential for accurate referring expression comprehension. Particularly noteworthy is our model's strong performance in zero target scenarios, achieving 47.9/86.0 Acc@0.25 with/without distractors compared to IPDN's 36.8/81.1 and MDIN's 22.2/67.2, demonstrating superior capability in distinguishing when no valid targets exist in the scene. Wile IPDN [6] already incorporates 2D features and serves as a strong multi-modal baseline, our approach also shows improvments when comparing with it across metrics. we further validate it through abalation studies and also observed some limitations in our design.

Is-RES also demonstrates strong performance in challenging multiple target scenarios, achieving 78.9 Acc@0.25 and 52.9 Acc@0.5, both representing the highest scores among all compared methods and highlighting our progressive fusion strategy's effectiveness in handling complex multi-object referring expressions.

On Scanrefer[5] benchmark, IS-RES achieves 59.9%/54.7%/49.9% on Acc@0.25/Acc@0.5/mIoU respectively. Compared to IPDN [6], IS-RES achieves identical performance on Acc@0.25 (59.9%), with modest improvements of 0.3% on Acc@0.5 and 0.2% on mIoU. Against MDIN[72], IS-RES demonstrates consistent improvements of 1.9% on Acc@0.25, 1.6% on Acc@0.5, and 1.6% on mIoU, validating the effectiveness of our hierarchical visual semantic decomposition approach.

The qualitative comparison is also listed at Figure 4.1. Some additional qualitative results are also presented, including both successful segmentation cases that highlight our method's capabilities in complex multi-target environments 4.2, as well as failure cases 4.3 that reveal current limitations and opportunities for future improvement.

TABLE 4.2. 3D-RES task results on ScanRefer. We used the mIoU and accuracy metrics based on the values provided in their respective papers.Acc@0.25 and Acc0.5 refer to the accuracy under IoU thresholds 0.25 and 0.5 respectively. Results marked with †indicate our own reproduction using the authors' published code.

Method	Venue	Unique (∼19%)			Multiple (∼81%)			**Overall**		
		0.25	0.5	mIoU	0.25	0.5	mIoU	**0.25**	**0.5**	**mIoU**
TGNN† [27]	AAAI2021	69.3	57.8	50.7	31.2	26.6	23.6	38.6	32.7	28.8
InstanceRefer [90]	ICCV2021	81.6	72.2	60.4	29.4	23.5	21.5	40.2	33.5	30.6
3DRefTR [70]	ECCV2024	89.6	77.0	-	52.3	43.7	-	57.9	48.7	41.2
X-RefSeg3D [56]	AAAI2024	-	-	-	-	-	-	40.3	33.8	29.9
3D-STMN [73]	AAAI2024	89.3	84.0	74.5	46.2	29.2	31.1	54.6	39.8	39.5
Reanson3D [26]	3DV2025	88.4	84.2	74.6	50.5	31.7	34.1	57.9	41.9	42.0
SegPoint [21]	ECCV2024	-	-	-	-	-	-	-	-	41.7
MCLN [57]	ECCV2024	89.6	78.2	-	**53.3**	45.9	-	58.7	50.7	44.7
RefMask3D [20]	ACMMM2024	89.6	84.7	-	48.1	40.8	-	55.9	49.2	44.9
MDIN [72]	ACMMM2024	91.0	87.2	76.7	50.1	44.9	41.4	58.0	53.1	48.3
IPDN† [6]	AAAI2025	**92.3**	88.0	78.1	52.0	46.4	42.8	**59.9**	54.4	49.7
Is-RES	-	92.2	**88.8**	**78.4**	52.2	**46.5**	**43.0**	**59.9**	**54.7**	**49.9**

4.3 Ablation Studies

We conduct ablation studies on our proposed two modules, hierarchical visual semantic decompositions and progressive multi-level fusion on Multi3dRefer[93] validation set. As shown in Table 4.3, we could see that both two modules contribute to modest improvement. When comparing our different feature configurations, we observe that the hierarchical visual semantic decomposition (VSD) 3.3 provides more substantial gains than the progressive multi-level fusion (MLF) module 3.4 when applied individually. Specifically, VSD alone achieves 72.0% Acc@0.25 compared to 71.5% with MLF alone, representing a 1.1% versus 0.6% improvement over the baseline respectively. The combination of both modules yields the optimal performance at 72.3% Acc@0.25, 53.4% Acc@0.5, and 53.5% mIoU, demonstrating that both strategies work synergistically to enhance 3D RES accuracy.

VSD	MLF	Acc@0.25	Acc@0.5	mIoU
✗	✗	70.9	50.5	51.5
✗	✓	71.5	51.5	52.3
✓	✗	72.0	52.0	52.9
✓	✓	72.3	53.4	53.5

TABLE 4.3. Ablation study on the proposed components on 3D-GRES setting. VSD stands for hierarchical visual semantic decompositions. And MLF stands for our progressive multi-level fusion module.

CHAPTER 5

Discussion

5.1 Conclusion

In this work, we present IS-RES, a *Unified Multi-Modal Approach* for 3D Referring Expression Segmentation that incorporates *Hierarchical Multi-Modal Collaborative Fusion* mechanisms. The Hierarchical Visual Semantic Decomposition component establishes an approach to extract visual representations at multiple semantic granularity. By leveraging SAM-guided instance segmentation and CLIP-based feature encoding, this component fundamentally transforms how visual semantics are captured and organized, moving beyond traditional uniform pixel treatment to semantically coherent instance-aware processing. The Progressive Multi-level Fusion component provides a systematic integration framework that achieves adaptive cross-modal alignment through carefully designed intra-modal collaboration and dynamic modal weighting mechanisms. This approach ensures that 2D visual semantics and 3D geometric information are effectively combined while maintaining semantic purity and preventing feature dilution across object boundaries. Experiments on ScanRefer and Multi3DRefer show that IS-RES achieves state-of-the-art results on both standard and *generalised* (multi-/zero-target) splits, with more gains in the zero-target scenarios.

5.2 Limitations and Future Directions

While IS-RES achieves substantial improvements on Multi3DRefer, particularly in zero-target scenarios, performance on ScanRefer remains comparable to IPDN, indicating architectural limitations that might constrain further gains. The multi-stage aggregation pipeline from 2D

pixel-level to superpoint-level representations may introduce cumulative information loss, while superpoint reliance inherently limits segmentation precision due to imperfect alignment with semantic boundaries.

Our current decomposition module's utilization of multi-view instance features may be sub-optimal, as multiple dimensional transformations required for cross-modal alignment might introduce additional noise. Our current high-level interaction approach, between 3D language-selected queries and visual features, might fail to fully exploit geometric correspondence between 3D superpoints and SAM-predicted 2D instance boundaries, suggesting more direct low-level interactions could better preserve semantic coherence.

Future work could explore direct geometric-semantic mapping between superpoints and SAM instance masks to maintain boundary integrity. Additionally, investigating hybrid representations beyond traditional superpoint clustering—such as adaptive multi-scale primitives—could enhance both computational efficiency and segmentation precision.

Bibliography

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny and Leonidas J. Guibas. 'ReferIt3D: Neural Listeners for Fine-Grained 3D Object Identification in Real-World Scenes'. In: *16th European Conference on Computer Vision (ECCV)*. 2020.

[2] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu and Chiew-Lan Tai. *TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection with Transformers*. 2022. arXiv: 2203.11496 [cs.CV]. URL: https://arxiv.org/abs/2203.11496.

[3] Eslam Bakr, Yasmeen Alsaedy and Mohamed Elhoseiny. 'Look around and refer: 2d synthetic semantics knowledge distillation for 3d visual grounding'. In: *Advances in neural information processing systems* 35 (2022), pp. 37146–37158.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov and Sergey Zagoruyko. *End-to-End Object Detection with Transformers*. 2020. arXiv: 2005.12872 [cs.CV].

[5] Dave Zhenyu Chen, Angel X Chang and Matthias Nießner. 'Scanrefer: 3d object localization in rgb-d scans using natural language'. In: *European conference on computer vision*. Springer. 2020, pp. 202–221.

[6] Qi Chen, Changli Wu, Jiayi Ji, Yiwei Ma, Danni Yang and Xiaoshuai Sun. *IPDN: Image-enhanced Prompt Decoding Network for 3D Referring Expression Segmentation*. 2025. arXiv: 2501.04995 [cs.CV]. URL: https://arxiv.org/abs/2501.04995.

[7] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao and Wenping Wang. *CLIP2Scene: Towards Label-efficient 3D Scene Understanding by CLIP*. 2023. arXiv: 2301.04926 [cs.CV].

[8] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid and Ivan Laptev. 'Language conditioned spatial relation reasoning for 3d object grounding'. In: *Advances in neural information processing systems* 35 (2022), pp. 20522–20535.

[9] Julian Chibane, Francis Engelmann, Tuan Anh Tran and Gerard Pons-Moll. *Box2Mask: Weakly Supervised 3D Semantic Instance Segmentation Using Bounding Boxes*. 2023. arXiv: 2206.01203 [cs.CV].

[10] Yong Xien Chng, Henry Zheng, Yizeng Han, Xuchong Qiu and Gao Huang. 'Mask grounding for referring image segmentation'. In: *CVPR*. 2024.

[11] Spconv Contributors. *Spconv: Spatially Sparse Convolution Library*. https://github.com/traveller59/spconv. 2022.

[12] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser and Matthias Nießner. 'Scannet: Richly-annotated 3d reconstructions of indoor scenes'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 5828–5839.

[13] Henghui Ding, Chang Liu, Suchen Wang and Xudong Jiang. 'Vision-language transformer and query generation for referring segmentation'. In: *ICCV*. 2021.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit and Neil Houlsby. 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale'. In: *International Conference on Learning Representations*. 2021.

[15] Chun Feng, Joy Hsu, Weiyu Liu and Jiajun Wu. 'Naturally Supervised 3D Visual Grounding with Language-Regularized Concept Learners'. In: *arXiv preprint arXiv:2404.19696* (2024).

[16] Estevão S. Gedraite and Murielle Hadad. 'Investigation on the effect of a Gaussian Blur in image filtering and segmentation'. In: *Proceedings ELMAR-2011*. 2011, pp. 393–396.

[17] Benjamin Graham, Martin Engelcke and Laurens Van Der Maaten. '3d semantic segmentation with submanifold sparse convolutional networks'. In: *CVPR*. 2018.

[18] Zoey Guo, Yiwen Tang, Ray Zhang, Dong Wang, Zhigang Wang, Bin Zhao and Xuelong Li. *ViewRefer: Grasp the Multi-view Knowledge for 3D Visual Grounding with GPT and Prototype Guidance*. 2023. arXiv: 2303.16894 [cs.CV]. URL: https://arxiv.org/abs/2303.16894.

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár and Ross Girshick. *Mask R-CNN*. 2018. arXiv: 1703.06870 [cs.CV]. URL: https://arxiv.org/abs/1703.06870.

[20] Shuting He and Henghui Ding. 'RefMask3D: Language-guided transformer for 3D referring segmentation'. In: *Proceedings of the 32nd ACM International Conference on Multimedia*. 2024, pp. 8316–8325.

[21] Shuting He, Henghui Ding, Xudong Jiang and Bihan Wen. 'SegPoint: Segment Any Point Cloud via Large Language Model'. In: *arXiv* (2024).

[22] Joy Hsu, Jiayuan Mao and Jiajun Wu. 'Ns3d: Neuro-symbolic grounding of 3d objects and relations'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 2614–2623.

[23] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell and Kate Saenko. 'Modeling Relationships in Referential Expressions with Compositional Modular Networks'. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 4418–4427. DOI: 10.1109/CVPR.2017.470. URL: https://doi.org/10.1109/CVPR.2017.470.

[24] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko and Trevor Darrell. 'Natural Language Object Retrieval'. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 4555–4564. DOI: 10.1109/CVPR.2016.493. URL: https://doi.org/10.1109/CVPR.2016.493.

[25] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang and Huchuan Lu. 'Bi-Directional Relationship Inferring Network for Referring Image Segmentation'. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 4423–4432. DOI: 10.1109/CVPR42600.2020.00448.

[26] Kuan-Chih Huang, Xiangtai Li, Lu Qi, Shuicheng Yan and Ming-Hsuan Yang. 'Reason3D: Searching and Reasoning 3D Segmentation via Large Language Model'. In: *arXiv* (2024).

[27] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen and Tyng-Luh Liu. 'Text-guided graph neural networks for referring 3d instance segmentation'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 2. 2021, pp. 1610–1618.

[28] Shijia Huang, Yilun Chen, Jiaya Jia and Liwei Wang. *Multi-View Transformer for 3D Visual Grounding*. 2022. arXiv: 2204.02174 [cs.CV]. URL: https://arxiv.org/abs/2204.02174.

[29] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta and Katerina Fragkiadaki. *Bottom Up Top Down Detection Transformers for Language Grounding in Images and Point Clouds*. 2022. arXiv: 2112.08879 [cs.CV].

[30] Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li and Tieniu Tan. *Locate then Segment: A Strong Pipeline for Referring Image Segmentation*. 2021. arXiv: 2103.16284 [cs.CV]. URL: https://arxiv.org/abs/2103.16284.

[31] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár and Ross Girshick. *Segment Anything*. 2023. arXiv: 2304.02643 [cs.CV]. URL: https://arxiv.org/abs/2304.02643.

[32] Maxim Kolodiazhnyi, Anna Vorontsova, Anton Konushin and Danila Rukhovich. *OneFormer3D: One Transformer for Unified Point Cloud Segmentation*. 2023. arXiv: 2311.14405 [cs.CV].

[33] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu and Jiaya Jia. 'Lisa: Reasoning segmentation via large language model'. In: *CVPR*. 2024.

[34] Loic Landrieu and Martin Simonovsky. 'Large-scale point cloud semantic segmentation with superpoint graphs'. In: *CVPR*. 2018.

[35] Feng Li, Hao Zhang, Huaizhe xu, Shilong Liu, Lei Zhang, Lionel M. Ni and Heung-Yeung Shum. *Mask DINO: Towards A Unified Transformer-based Framework for Object Detection and Segmentation*. 2022. arXiv: 2206.02777 [cs.CV]. URL: https://arxiv.org/abs/2206.02777.

[36] Xin Li, Tao Ma, Yuenan Hou, Botian Shi, Yuchen Yang, Youquan Liu, Xingjiao Wu, Qin Chen, Yikang Li, Yu Qiao and Liang He. *LoGoNet: Towards Accurate 3D Object Detection with Local-to-Global Cross-Modal Fusion*. 2023. arXiv: 2303.03595 [cs.CV]. URL: https://arxiv.org/abs/2303.03595.

[37] Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun and Jiaya Jia. *Unifying Voxel-based Representation with Transformer for 3D Object Detection*. 2022. arXiv: 2206.00630 [cs.CV]. URL: https://arxiv.org/abs/2206.00630.

[38] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang and Zhi Tang. *BEVFusion: A Simple and Robust LiDAR-Camera Fusion Framework*. 2022. arXiv: 2205.13790 [cs.CV]. URL: https://arxiv.org/abs/2205.13790.

[39] Haojia Lin, Yongdong Luo, Xiawu Zheng, Lijiang Li, Fei Chao, Taisong Jin, Donghao Luo, Chengjie Wang, Yan Wang and Liujuan Cao. 'A Unified Framework for 3D Point Cloud Visual Grounding'. In: *arXiv preprint arXiv:2308.11887* (2023).

[40] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He and Piotr Dollár. *Focal Loss for Dense Object Detection*. 2018. arXiv: 1708.02002 [cs.CV]. URL: https://arxiv.org/abs/1708.02002.

[41] Chang Liu, Henghui Ding and Xudong Jiang. *GRES: Generalized Referring Expression Segmentation*. 2023. arXiv: 2306.00968 [cs.CV].

[42] Chang Liu, Xudong Jiang and Henghui Ding. 'Instance-specific feature propagation for referring segmentation'. In: *IEEE Transactions on Multimedia* (2022).

[43] Jingyu Liu, Liang Wang and Ming-Hsuan Yang. 'Referring Expression Generation and Comprehension via Attributes'. In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 4866–4874. DOI: 10.1109/ICCV.2017.520. URL: https://doi.org/10.1109/ICCV.2017.520.

[44] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov. 'Roberta: A robustly optimized bert pretraining approach'. In: *arXiv preprint arXiv:1907.11692* (2019).

[45] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus and Song Han. *BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation*. 2024. arXiv: 2205.13542 [cs.CV]. URL: https://arxiv.org/abs/2205.13542.

[46] Ilya Loshchilov and Frank Hutter. *Decoupled Weight Decay Regularization*. 2019. arXiv: 1711.05101 [cs.LG]. URL: https://arxiv.org/abs/1711.05101.

[47] Junyu Luo, Jiahui Fu, Xianghao Kong, Chen Gao, Haibing Ren, Hao Shen, Huaxia Xia and Si Liu. '3d-sps: Single-stage 3d visual grounding via referred point progressive selection'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 16454–16463.

[48] Ruotian Luo and Gregory Shakhnarovich. 'Comprehension-Guided Referring Expressions'. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 3125–3134. DOI: 10.1109/CVPR.2017.333. URL: https://doi.org/10.1109/CVPR.2017.333.

[49] Guofeng Mei, Luigi Riz, Yiming Wang and Fabio Poiesi. *Vocabulary-Free 3D Instance Segmentation with Vision and Language Assistant*. 2025. arXiv: 2408.10652 [cs.CV]. URL: https://arxiv.org/abs/2408.10652.

[50] Carsten Moenning and Neil A Dodgson. *Fast marching farthest point sampling*. Tech. rep. University of Cambridge, Computer Laboratory, 2003.

[51] Li Muchen and Sigal Leonid. 'Referring Transformer: A One-step Approach to Multi-task Visual Grounding'. In: *Thirty-Fifth Conference on Neural Information Processing Systems*. 2021.

[52] Phuc D. A. Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham and Khoi Nguyen. *Open3DIS: Open-Vocabulary 3D Instance Segmentation with 2D Mask Guidance*. 2024. arXiv: 2312.10671 [cs.CV]. URL: https://arxiv.org/abs/2312.10671.

[53] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser et al. 'Openscene: 3d scene understanding with open vocabularies'. In: *CVPR*. 2023.

[54] Xidong Peng, Runnan Chen, Feng Qiao, Lingdong Kong, Youquan Liu, Yujing Sun, Tai Wang, Xinge Zhu and Yuexin Ma. *Learning to Adapt SAM for Segmenting Cross-domain Point Clouds*. 2024. arXiv: 2310.08820 [cs.CV]. URL: https://arxiv.org/abs/2310.08820.

[55] Charles R. Qi, Or Litany, Kaiming He and Leonidas J. Guibas. *Deep Hough Voting for 3D Object Detection in Point Clouds*. 2019. arXiv: 1904.09664 [cs.CV]. URL: https://arxiv.org/abs/1904.09664.

[56] Zhipeng Qian, Yiwei Ma, Jiayi Ji and Xiaoshuai Sun. 'X-RefSeg3D: Enhancing Referring 3D Instance Segmentation via Structured Cross-Modal Graph Neural Networks'. In: *AAAI*. 2024.

[57] Zhipeng Qian, Yiwei Ma, Zhekai Lin, Jiayi Ji, Xiawu Zheng, Xiaoshuai Sun and Rongrong Ji. 'Multi-branch Collaborative Learning Network for 3D Visual Grounding'. In: *arXiv* (2024).

[58] 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. arXiv: 2103.00020 [cs.CV]. URL: https://arxiv.org/abs/2103.00020.

[59] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár and Christoph Feichtenhofer. *SAM 2: Segment Anything in Images and Videos*. 2024. arXiv: 2408.00714 [cs.CV]. URL: https://arxiv.org/abs/2408.00714.

[60] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei and Christopher D Manning. 'Generating semantically precise scene graphs from textual descriptions

for improved image retrieval'. In: *Proceedings of the fourth workshop on vision and language*. 2015, pp. 70–80.

[61] Nisarg A Shah, Vibashan VS and Vishal M Patel. 'LQMFormer: Language-aware Query Mask Transformer for Referring Image Segmentation'. In: *CVPR*. 2024.

[62] Chao Shang, Zichen Song, Heqian Qiu, Lanxiao Wang, Fanman Meng and Hongliang Li. 'Prompt-Driven Referring Image Segmentation with Instance Contrasting'. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 4124–4134. DOI: 10.1109/CVPR52733.2024.00395.

[63] Jiahao Sun, Chunmei Qing, Junpeng Tan and Xiangmin Xu. 'Superpoint Transformer for 3D Scene Instance Segmentation'. In: *arXiv preprint arXiv:2211.15766* (2022).

[64] Naoya Takahashi and Yuki Mitsufuji. 'D3net: Densely connected multidilated densenet for music source separation'. In: *arXiv preprint arXiv:2010.01733* (2020).

[65] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari and Francis Engelmann. *OpenMask3D: Open-Vocabulary 3D Instance Segmentation*. 2023. arXiv: 2306.13631 [cs.CV]. URL: https://arxiv.org/abs/2306.13631.

[66] Jiajin Tang, Ge Zheng, Cheng Shi and Sibei Yang. 'Contrastive Grouping With Transformer for Referring Image Segmentation'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2023, pp. 23570–23580.

[67] Ozan Unal, Christos Sakaridis, Suman Saha and Luc Van Gool. 'Four Ways to Improve Verbo-visual Fusion for Dense 3D Visual Grounding'. In: *European Conference on Computer Vision (ECCV)*. Oct. 2024.

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser and Illia Polosukhin. 'Attention is All you Need'. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[69] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong and Tongliang Liu. 'Cris: Clip-driven referring image segmentation'. In: *CVPR*. 2022.

[70] Zhenyu Wang, Yali Li, Taichi Liu, Hengshuang Zhao and Shengjin Wang. 'OV-Uni3DETR: Towards Unified Open-Vocabulary 3D Object Detection via Cycle-Modality Propagation'. In: *ECCV2024* (2024).

[71] Changli Wu, Qi Chen, Jiayi Ji, Haowei Wang, Yiwei Ma, You Huang, Gen Luo, Hao Fei, Xiaoshuai Sun and Rongrong Ji. *RG-SAN: Rule-Guided Spatial Awareness Network for End-to-End 3D Referring Expression Segmentation*. 2024. arXiv: 2412.02402 [cs.CV]. URL: https://arxiv.org/abs/2412.02402.

[72] Changli Wu, Yihang Liu, Jiayi Ji, Yiwei Ma, Haowei Wang, Gen Luo, Henghui Ding, Xiaoshuai Sun and Rongrong Ji. '3d-gres: Generalized 3d referring expression segmentation'. In: *Proceedings of the 32nd ACM International Conference on Multimedia*. 2024, pp. 7852–7861.

[73] Changli Wu, Yiwei Ma, Qi Chen, Haowei Wang, Gen Luo, Jiayi Ji and Xiaoshuai Sun. '3D-STMN: Dependency-Driven Superpoint-Text Matching Network for End-to-End 3D Referring Expression Segmentation'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 6. 2024, pp. 5940–5948.

[74] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun and Wei-Ying Ma. 'Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations'. In: *CVPR*. 2019, pp. 6609–6618.

[75] Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng and Jian Zhang. 'EDA: Explicit Text-Decoupling and Dense Alignment for 3D Visual Grounding'. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023. DOI: 10.1109/cvpr52729.2023.01843. URL: http://dx.doi.org/10.1109/CVPR52729.2023.01843.

[76] Can Xu, Yuehui Han, Rui Xu, Le Hui, Jin Xie and Jian Yang. 'Multi Attributes Interactions Matters for 3D Visual Grounding'. In: *CVPR*. 2024.

[77] Shaoqing Xu, Dingfu Zhou, Jin Fang, Junbo Yin, Zhou Bin and Liangjun Zhang. *FusionPainting: Multimodal Fusion with Adaptive Attention for 3D Object Detection*.

2021. arXiv: 2106.12449 [cs.CV]. URL: https://arxiv.org/abs/2106.12449.

[78] Wei Xu, Chunsheng Shi, Sifan Tu, Xin Zhou, Dingkang Liang and Xiang Bai. 'A Unified Framework for 3D Scene Understanding'. In: *arXiv* (2024).

[79] Zunnan Xu, Zhihong Chen, Yong Zhang, Yibing Song, Xiang Wan and Guanbin Li. 'Bridging Vision and Language Encoders: Parameter-Efficient Tuning for Referring Image Segmentation'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2023, pp. 17503–17512.

[80] Danni Yang, Jiayi Ji, Yiwei Ma, Tianyu Guo, Haowei Wang, Xiaoshuai Sun and Rongrong Ji. 'SAM as the Guide: Mastering Pseudo-Label Refinement in Semi-Supervised Referring Expression Segmentation'. In: *arXiv* (2024).

[81] Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu and Li Zhang. *DeepInteraction: 3D Object Detection via Modality Interaction*. 2022. arXiv: 2208.11112 [cs.CV]. URL: https://arxiv.org/abs/2208.11112.

[82] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao and Philip HS Torr. 'Lavt: Language-aware vision transformer for referring image segmentation'. In: *CVPR*. 2022.

[83] Linwei Ye, Mrigank Rochan, Zhi Liu and Yang Wang. *Cross-Modal Self-Attention Network for Referring Image Segmentation*. 2019. arXiv: 1904.04745 [cs.CV]. URL: https://arxiv.org/abs/1904.04745.

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

[85] Junbo Yin, Jianbing Shen, Runnan Chen, Wei Li, Ruigang Yang, Pascal Frossard and Wenguan Wang. *IS-Fusion: Instance-Scene Collaborative Fusion for Multimodal 3D Object Detection*. 2024. arXiv: 2403.15241 [cs.CV]. URL: https://arxiv.org/abs/2403.15241.

[86] Tianwei Yin, Xingyi Zhou and Philipp Krähenbühl. *Multimodal Virtual Point 3D Detection*. 2021. arXiv: 2111.06881 [cs.CV]. URL: https://arxiv.org/abs/2111.06881.

[87] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal and Tamara L. Berg. *MAttNet: Modular Attention Network for Referring Expression Comprehension*. 2018. arXiv: 1801.08186 [cs.CV]. URL: https://arxiv.org/abs/1801.08186.

[88] Licheng Yu, Hao Tan, Mohit Bansal and Tamara L. Berg. 'A Joint Speaker-Listener-Reinforcer Model for Referring Expressions'. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 3521–3529. DOI: 10.1109/CVPR.2017.375. URL: https://doi.org/10.1109/CVPR.2017.375.

[89] Qingtao Yu, Heming Du, Chen Liu and Xin Yu. 'When 3D Bounding-Box Meets SAM: Point Cloud Instance Segmentation with Weak-and-Noisy Supervision'. In: *WACV*. 2024.

[90] Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Sheng Wang, Zhen Li and Shuguang Cui. 'Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring'. In: *ICCV*. 2021.

[91] Haomeng Zhang, Chiao-An Yang and Raymond A. Yeh. *Multi-Object 3D Grounding with Dynamic Modules and Language-Informed Spatial Attention*. 2024. arXiv: 2410.22306 [cs.CV]. URL: https://arxiv.org/abs/2410.22306.

[92] Junbo Zhang, Runpei Dong and Kaisheng Ma. 'Clip-fo3d: Learning free open-world 3d scene representations from 2d dense clip'. In: *ICCV*. 2023.

[93] Yiming Zhang, ZeMing Gong and Angel X Chang. 'Multi3drefer: Grounding text description to multiple 3d objects'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 15225–15236.

[94] Zicheng Zhang, Yi Zhu, Jianzhuang Liu, Xiaodan Liang and Wei Ke. *CoupAlign: Coupling Word-Pixel with Sentence-Mask Alignments for Referring Image Segmentation*. 2022. arXiv: 2212.01769 [cs.CV]. URL: https://arxiv.org/abs/2212.01769.