IntentFuse: Language-Guided 3D Scene Understanding via Prompt Filtering and Fusion

Ahalya Ravendran, Madhawa Perera, Feng Xu, Lars Petersson, Dadong Wang, Xun Li
CSIRO Data61
Australia
{ahalya.ravendran, madhawa.perera, feng.xu, lars.petersson, dadong.wang, xun.li}@data61.csiro.au

***Abstract*—Affordances—the actions enabled by objects—are fundamental for meaningful interaction in 3D environments. While language models can infer affordances from abstract prompts, grounding them in visual scenes typically requires extensive supervision. We propose IntentFuse, a lightweight middleware that integrates a compact language model with a pretrained Language Embedded Radiance Field (LERF) to ground affordance reasoning directly in 3D scenes. Our method reformulates free-form queries into structured CLIP-aligned prompts, extracting key roles and filtering negations before injecting them into LERF. Experiments on descriptive, affordance, and negation benchmarks show clear improvements, achieving 95% accuracy on affordance queries and 88% on negation handling over baseline LERF. IntentFuse enables intuitive, training-free, plug-and-play affordance grounding with applications in real-time robotic instruction following and AR/VR scene exploration, paving the way for more natural and interpretable human–environment interaction.**

***Index Terms*—affordance grounding, scene understanding, language-guided visual search, neural radiance fields, prompt filtering**

I. Introduction

As intelligent agents transition from controlled laboratory environments to real-world settings, the ability to comprehend and act upon natural human language becomes increasingly critical. In collaborative robotic domains, linguistic directives are often abstract, non-specific, or heavily context-dependent. For example, a human operator may express an under-specified utterance such as "I am feeling hungry," rather than issuing a precise instruction like "proceed to the left, identify a red apple within the environment, and deliver it to the operator." To function robustly under these conditions, agents must infer user intent, disambiguate ambiguous language, and localize semantically relevant entities within 3D scenes based on high-level verbal cues.

Recent language-grounded 3D representations [1]–[4] integrate CLIP embeddings to map text prompts into spatial relevance fields, enabling open-vocabulary search in reconstructed scenes. While effective, these models inherit caption-level biases from CLIP training and often struggle with affordances ("something to sit on"), negations ("not a chair"), and user intentions, posing challenges for task-directed grounding in robotics.

In this paper, we introduce IntentFuse, a lightweight, plug-and-play middleware designed to bridge the gap between free-form natural language and CLIP-guided 3D visual grounding systems. As agents transition into real-world environments, understanding high-level user intent, including affordances, negations, and descriptive cues, remains challenging for existing language-embedded 3D models that operate on unstructured prompts. IntentFuse addresses this by applying structured prompting from a pretrained language model, reformulating free-form queries into CLIP-aligned representations that better reflect user intention. The pipeline comprises two core components: a Query Evaluator, which extracts semantic roles (object, affordance, attribute, intention, negation), and an optional Context Provider, which incorporates environment-specific priors when available. Without requiring additional training or architectural modifications, IntentFuse enhances relevance mapping in existing 3D grounding models, producing sharper, context-aware activations as illustrated in

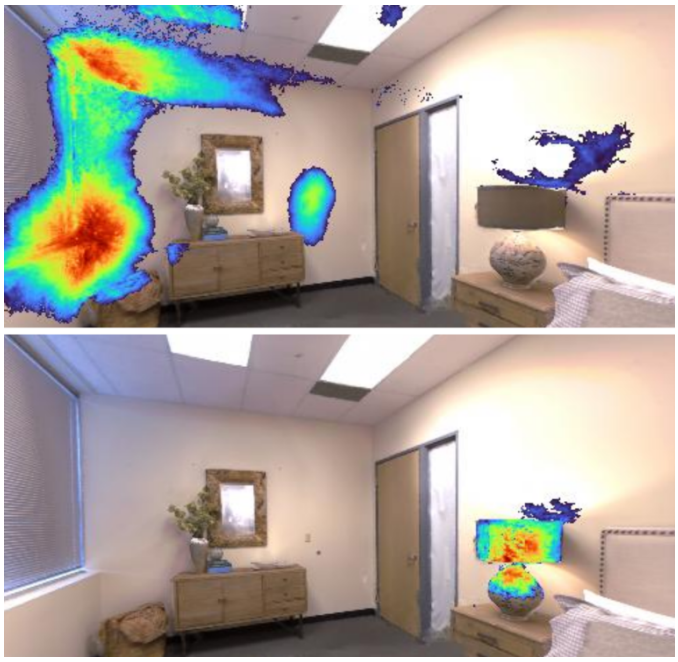

Fig. 1. Intent-aware Relevance Mapping in 3D Scenes. Given the abstract query "something to light up my desk at night", LERF (top) activates irrelevant bright surfaces such as the wall. In contrast, IntentFuse (bottom) accurately localizes the desk lamp, demonstrating improved intent-aware grounding in 3D scenes.

Fig. 1. This architecture-agnostic design opens new possibilities for intent-aware scene understanding in robotics, enabling more natural and effective human-agent collaboration. Code and dataset are available at https://collaborative-work-space.github.io/intentfuse/. Our main contributions are:

- **Intent-Guided Prompt Filtering.** We introduce a structured prompting strategy that extracts semantic roles such as object, affordance, attribute, intention, and negation, from free-form queries using a pretrained LLM.
- **Plug-and-Play Integration.** Operating as a training-free post-processing step, our module integrates seamlessly into existing language-embedded 3D vision systems and enhances their ability to construct and interpret scene graphs from nuanced language inputs.
- **Improved Contextual Grounding.** We demonstrate superior performance over a baseline LERF on complex, intent-driven queries across twelve diverse indoor scenes, six synthetic environments from Replica (chosen for varied layouts and object densities) and six real-world scenes captured with handheld cameras under different lighting conditions, enabling closer alignment with human intent in both simulated and real-world settings.

II. Related Work

Neural Scene Rendering. Neural Radiance Fields (NeRFs) [5] and their accelerated variants [6], [7] have emerged as foundational methods for high-fidelity 3D scene reconstruction. More recently, Gaussian Splatting [8] has offered a real-time alternative with competitive visual quality. These models encode rich geometric and photometric detail, but lack semantic understanding or language grounding, limiting their utility in interactive or task-driven settings in robotics.

Language-Embedded Scene Reconstruction. To bridge language and vision in 3D, recent methods incorporate CLIP [9] into reconstruction pipelines. LERF [1], LangSplat [3], 4DLangSplat [4], and IGS2GS [2] extend neural and splatting-based representations to support open-vocabulary queries. These approaches construct 3D language fields either via volume rendering of CLIP features, Gaussian-based embeddings, or temporal extensions. However, they rely on caption-level supervision that often overlooks complex linguistic structures involving affordances, negation, and intention. As a result, they may over-activate irrelevant regions or misalign with user goals in task-oriented scenarios.

Affordance and Scene Graph Understanding. Scene graphs have gained growing interest in the domain of scene understanding in recent years, finding wide applications in downstream tasks such as visual question answering [10], robotic tasks [11], and intelligent agents [12]. Affordances and scene graphs are critical for enabling robotic agents to reason about object functionality and spatial relationships [13]–[17]. While models like LERF and LangSplat generate dense relevance maps from language, they do not explicitly model scene graphs. Extracting graph nodes and edges from volumetric relevancy outputs—especially without segmentation masks—relies heavily on raw CLIP activations, which are insufficient for understanding affordances, negation, or intent. Consequently, scene graph construction from these representations often fails to capture task-relevant semantics. We address this gap with a structured, intent-aware prompting layer that produces CLIP-aligned queries better suited for scene-level graph reasoning.

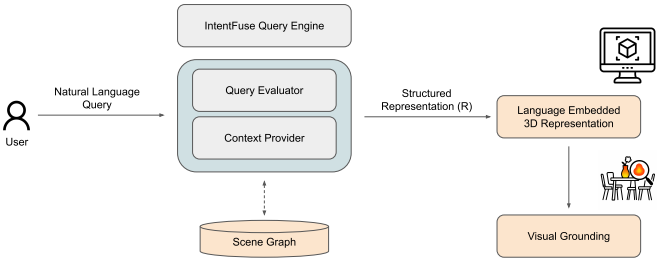

Fig. 2. IntentFuse Query Engine overview. The Query Evaluator processes natural language input, analyzing its semantic structure and extracting key roles for the LERF engine. The Context Provider optionally leverages environment priors (e.g., scene graphs or object dictionaries) to resolve ambiguities and semantically enrich the query. The resulting structured representation R is then forwarded to the LERF Query Engine for precise 3D visual grounding.

Virtual Simulators vs. Real-World Grounding. Many approaches to intent grounding and affordance learning are evaluated in simulation environments such as AI2-THOR [18] or Habitat [19], where object affordances are predefined and visual variability is limited. In contrast, we focus on grounding free-form queries in reconstructed 3D scenes from both synthetic (Replica [20]) and real-world handheld captures, presenting challenges like unstructured environments and ambiguous instructions and underscoring the need for more generalizable, language-aware systems.

These gaps highlight the need for structured prompting approaches that can bridge affordances, negation, and intent, enabling more reliable 3D grounding in both synthetic and real-world environments.

III. Method

IntentFuse Query Engine functions as a middleware that transforms free-form user queries—including negation, affordance, descriptive, and direct object requests—via a pretrained LLM into CLIP-aligned prompts. These structured prompts are then consumed by language-embedded scene representation, be it the LERF Query Engine or any CLIP-based grounding pipeline over images, point clouds, or volumetric data to produce dense, view-consistent relevancy maps for 3D scenes without additional training or architectural changes, as shown in Fig. 2

A. Query Evaluator

Given a user-provided natural language query p, we invoke the `gemma3:1b` text input model [21], [22] to restructure it into a form s suitable for querying 3D environments using the CLIP model leveraged by LERF [1]. Consistent with the design principles outlined in Li et al. [23], the query evaluator interprets the semantics of the user query and generates a structured representation R. This representation encodes

TABLE I
QUERY RESTRUCTURING EXAMPLES ILLUSTRATING SEMANTIC ROLE EXTRACTION (EVALUATOR) AND CONTEXT-DRIVEN REFINEMENT (PROVIDER).

User Query	Evaluator Output	Context Provider
Something showing time, unlike rotating dial clock.	*Negation*: rotating dial clock Object: none *Affordance*: something showing time *Description*: none	*Object*: digital dial clock *Affordance*: something showing time
Something wooden, unlike a soft toy.	*Negation*: soft toy Object: none *Affordance*: none *Description*: something wooden	*Object*: coffee table *Description*: something wooden
Deep blue armchair with smooth texture for comfortable seating.	*Negation*: none *Object*: armchair *Affordance*: comfortable seating *Description*: deep blue armchair with smooth texture	*Object*: armchair *Affordance*: comfortable seating *Description*: deep blue armchair with smooth texture

constructs such as *negation*, *affordance*, *object of interest*, and *description*, thereby aligning natural language input with the scene representations used in downstream modules. Table I demonstrates examples of how user queries are restructured by the query evaluator.

B. Context Provider

As an optional component that relies on the availability of scene representation, the context provider further refines the structured representation R by incorporating environmental priors, thereby enhancing grounding precision. For example, in a warehouse with a known inventory, the environment can be represented as a scene graph or simple object list E_r. In our implementation, following the context-aware augmentation strategy in Li et al. [23], E_r is instantiated as an object list automatically extracted from the reconstructed scene. Given an under-specified query such as "something to sit on," the query evaluator may produce an affordance-centric representation (object: none, affordance: "something to sit on"). The context provider then augments R by populating the object slot with contextually appropriate candidates from E_r (e.g., chair, bench, sofa), enabling the LERF engine to perform more accurate, context-aware 3D grounding. The resulting structured representation R is then passed to the LERF query engine, which generates dense, view-consistent relevancy maps for visual grounding. This streamlined pipeline enables high-precision affordance and intent understanding in 3D scenes without additional training or architectural changes.

C. Prompt Compilation for CLIP Grounding

From the extracted fields in R, we compile a grounding string by combining the *negation, affordance, object,* and *description* components. Negations are treated separately as suppressive cues, allowing the grounding pipeline to actively avoid irrelevant regions. This structured representation can be injected into any CLIP-guided visual grounding system by aligning positive prompts with target semantics and negative prompts with distractors. In our implementation with LERF, we pass these as `positives([clip prompt])` and `negatives(negations)` to guide relevance mapping toward semantically appropriate areas while suppressing distractors

D. Relevancy Map Visualisation

We integrate with the Nerfstudio evaluation pipeline to visualize dense relevancy maps without any additional training. For each camera pose, we cast rays using the known intrinsic and extrinsic parameters and run the LERF model to produce both RGB renderings and CLIP-based relevancy outputs. We then save these per-view relevancy maps for qualitative visualization and quantitative analysis. Our visualization pipeline handles both high-fidelity synthetic indoor scans and real-world reconstructions from RGB video, showcasing its generality across simulated and real 3D environments.

IV. RESULTS

We evaluate our system on both synthetic and real-world indoor 3D environments, using the Replica dataset and a custom indoor scan captured with a handheld RGB camera as shown in Fig. 3. We evaluate the impact of IntentFuse enabled query on CLIP-based 3D visual grounding using both qualitative visualizations and quantitative relevance metrics.

A. Datasets

Synthetic Dataset: We use the Replica dataset, a high-fidelity 3D scan collection of office and room interiors widely adopted for semantic scene understanding and embodied AI. Each scene provides 900 high-resolution synthetic images, offering a standard benchmark for evaluating 3D visual grounding in controlled indoor environments.

Real-World Dataset: We collect multi-view RGB footage of indoor environments using an iPhone 16 Pro (48 MP Main Fusion camera). For each scene, 30 keyframes are extracted, and camera poses are reconstructed with COLMAP

Fig. 3. Example images from our dataset. Synthetic Replica scenes (top) depict office and room variations with diverse object layouts. Real-world captures (bottom) showcase indoor environments with rich variation in texture, color, and geometry

to generate LERF-compatible scenes. This setup supports real-time rendering and post-processing, enabling evaluation under realistic, cluttered conditions.

B. Query Generation

We generate three types of language queries—descriptive, affordance, and negation—for 3D scene objects using OpenAI's GPT-4o model. To guide generation, the prompt includes definitions for each query type, multiple annotated examples, and explicit instructions to produce concise outputs. Example descriptive queries include "These are table legs and they are underneath the table" and "This is a side table under a lamp." For affordance queries, examples such as "Something I can open with my keys" and "Something to dispose of wastepaper in" illustrate functional reasoning. Negation queries are exemplified by "Something small, unlike a couch" and "Something rigid, unlike a cushion." We set the temperature parameter to 0.7, balancing creativity with consistency. This procedure yields 10 diverse queries per type for each room, enabling robust evaluation of IntentFuse's ability to generalize across reasoning categories.

C. Qualitative Evaluation

We evaluate IntentFuse qualitatively by comparing it with the baseline LERF [1], focusing on how IntentFuse enabled querying enhances 3D visual grounding across a range of query types: object-only, descriptive, affordance-aware, and negation-based. These categories reflect increasing levels of reasoning complexity and align with the structured semantic roles extracted by our Query Evaluator (see Fig. 2). While IntentFuse extracts additional spatial and relational fields, this evaluation emphasizes object-centric grounding as our datasets lack ground truth scene graph annotations for relational evaluation.

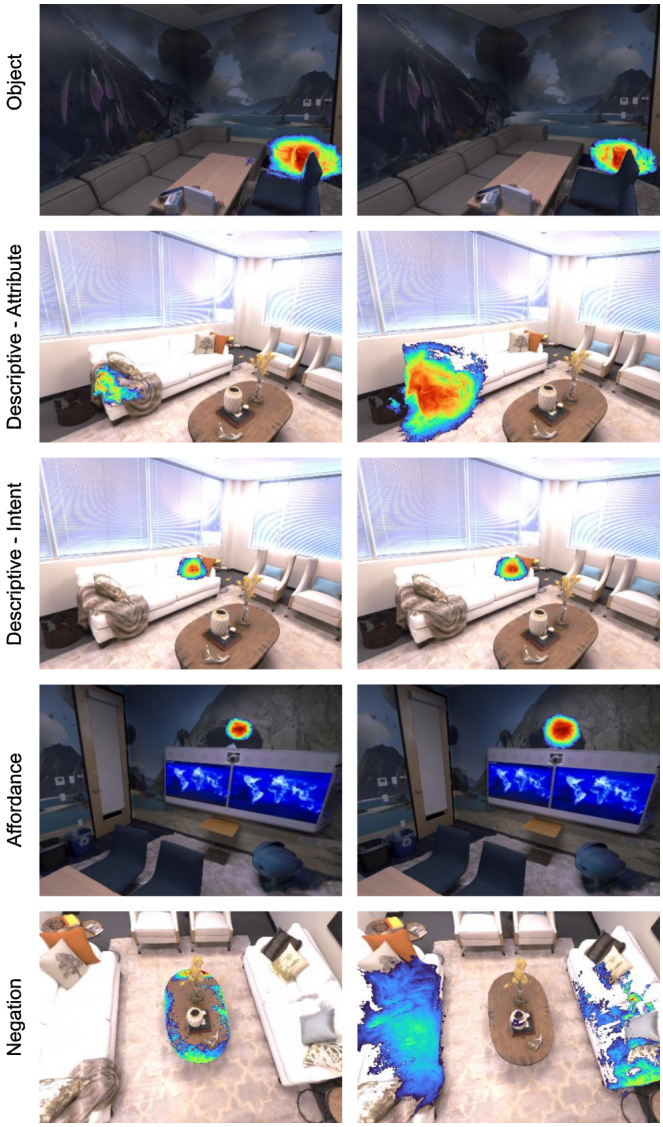

Fig. 4. Qualitative comparison between LERF (left) and IntentFuse (right) across grounding scenarios. Top: object-only query ("garbage bin"). Middle: descriptive queries ("Decorative pillow with tree silhouette in cream and brown", "This plush throw blanket is rich taupe and muted brown"). Bottom: affordance ("something to tell time") and negation ("something soft, unlike a solid table"). IntentFuse yields sharper, more localized activations across all query types.

Fig. 4 illustrates representative comparisons in both Replica synthetic environments and real-world reconstructed indoor scenes. Across all query types, IntentFuse consistently produces sharper, more localized activation maps, with reduced background noise and clearer object boundaries relative to LERF.

Object-Only Queries. For simple object references (e.g., "garbage bin"), both methods localize targets correctly. However, IntentFuse generates more compact and precise activations (Fig. 4, top row), highlighting its advantage even in straightforward scenarios by filtering extraneous activations via structured prompting.

Descriptive Queries (Intent + Attribute). Descriptive

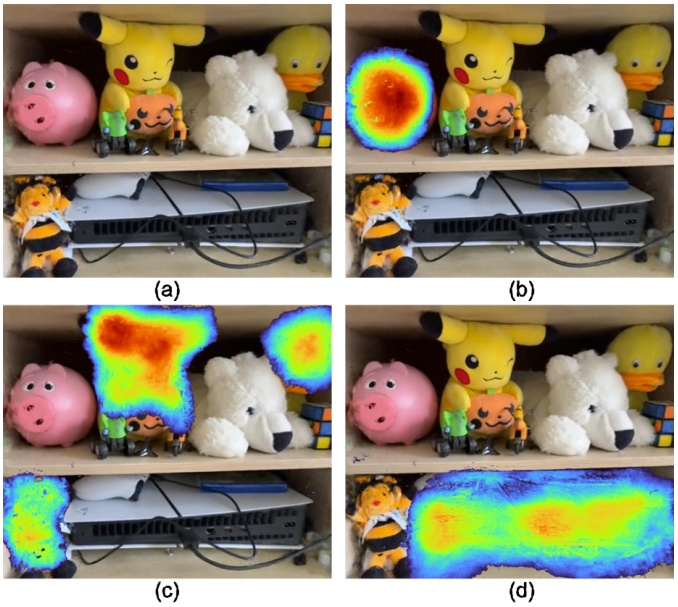

Fig. 5. Descriptive grounding results using IntentFuse in real-world scenes. (a) Original image. (b) Abstract query ("farm companion") activates the pig toy. (c) Attribute query ("yellow toy") highlights yellow plushes. (d) Intent query ("this is used to play games") isolates a game console

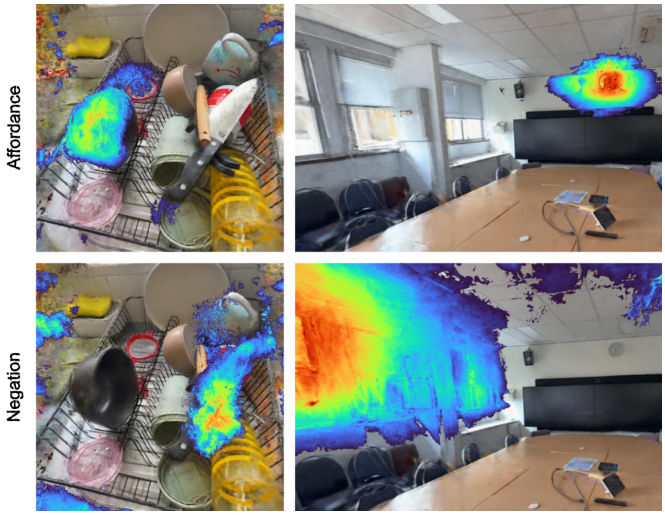

Fig. 6. Grounding affordance and negation queries in real-world kitchen and office scenes. Top: affordances ("something to hold fruits", "something to tell time"). Bottom: negations ("something to cut fruits unlike spoon", "something to allow natural light unlike bulbs"). IntentFuse distinctly highlights functional targets and suppresses distractors.

queries combine attributes and high-level intent, such as "Decorative pillow with tree silhouette in cream and brown" and "This plush throw blanket is rich taupe and muted brown." Here, IntentFuse disambiguates subtle linguistic cues and highlights the relevant objects (matching pillow and throw), whereas LERF's responses remain overly diffuse (Fig. 4, middle row). This reflects IntentFuse's ability to leverage both attribute and intent fields extracted in the Query Evaluator stage.

Affordance-Aware Queries. Grounding affordances (e.g., "something to hold fruits", "something to tell time") requires reasoning about object functionality, not just appearance. IntentFuse expands such prompts into role-aware grounding sentences using context provider and identifies relevant objects (black bowl in dish rack, wall clock) with minimal noise. In contrast, LERF's activations often bleed onto nearby but functionally irrelevant items (Fig. 4, bottom row).

Negation-Based Queries. Negation-aware grounding, such as "something to cut fruits unlike spoon" or "something to allow natural light unlike bulbs", challenges models to explicitly suppress distractors. By incorporating negation fields, IntentFuse sharply activates on the knife and window while suppressing irrelevant items like spoons or bulbs (Fig. 4, bottom row). LERF, lacking explicit negation handling, exhibits broader and less discriminative activations.

Real-World Captured Scenes. Fig. 5 shows descriptive queries (`farm companion`, `yellow toy`, `this is used to play games`), with crisp localization of the pig toy, yellow plushes, and game console. These results demonstrate that IntentFuse transfers robustly to real-world scenes—outperforming broadly activated maps of LeRF by leveraging affordance, attribute, and intent cues.

We further validate IntentFuse on real-world handheld RGB captures, focusing on cluttered kitchen and office environments reconstructed from handheld captured data. As shown in Fig. 6, these scenes present significant challenges for visual grounding, including occlusions, variable lighting, and dense object arrangements not present in synthetic datasets. Despite these complexities, IntentFuse demonstrates robust performance across both affordance ("something to hold fruits", "something to tell time") and negation-based queries ("something to allow natural light unlike bulbs", "something to cut fruits unlike spoon").

For affordance queries, IntentFuse's structured representation enables the model to identify functionally relevant targets (e.g., a black bowl in the dish rack or a wall-mounted clock) while suppressing visually salient yet functionally irrelevant distractors such as nearby utensils or decorative items. In negation-based prompts, IntentFuse effectively filters out specified distractors by leveraging its negation field during structured prompt compilation. This prevents the propagation of negated terms into CLIP, a common failure mode in baseline LERF where irrelevant objects (e.g., bulbs or spoons) receive high activation.

These results highlight the plug-and-play nature of IntentFuse within CLIP-guided 3D pipelines: no retraining or architectural modifications were required to adapt to real-world data. The Query Evaluator decomposes under-specified free-form queries into semantic roles even in visually noisy conditions, while the Context Provider (when available) resolves ambiguities using scene priors. Together, they enable IntentFuse to generalize across domains—bridging the gap between abstract language intent and concrete scene-level relevance in both synthetic and real-world 3D environments.

TABLE II
QUANTITATIVE EVALUATION ACROSS QUERY TYPES ON REPLICA DATASET. DESCRIPTIVE QUERIES MEASURED BY PRECISION/RECALL; AFFORDANCE AND NEGATION QUERIES BY SUCCESS RATE. INTENTFUSE OUTPERFORMS LERF ACROSS ALL CATEGORIES.

Methods	Descriptive		Affordance	Negation
	Precision (%)	Recall (%)	Score (%)	Score (%)
LeRF	0.71	0.88	0.83	0.43
Ours	**0.93**	**0.92**	**0.95**	**0.88**

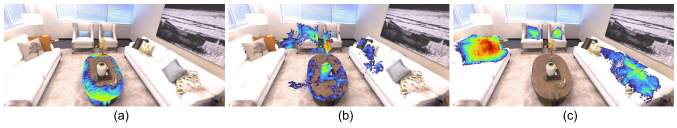

Fig. 7. Representative failure cases of IntentFuse. Left: Negation query ("something soft unlike table") shows LERF activating on a table, while IntentFuse suppresses it but highlights a few soft materials (blankets, cushions). Right: Abstract query ("something soft") where the intended target (pillow) is present with environment priors. Tight coupling of the context provider with user intent modeling could address intention ambiguities in future iterations

D. Quantitative Evaluation

We quantitatively evaluate on Replica across four query types: descriptive (attributes + intent), affordance, and negation-based prompts. We report precision and recall for descriptive queries and use success rates to evaluate functional and negation-based prompts, reflecting the system's ability to align language cues with relevant 3D regions.

As shown in Table II, IntentFuse achieves 0.93 precision and 0.92 recall on descriptive queries, outperforming LERF baseline. This 30% relative precision gain reflects how structured semantic roles (attribute, intent, object) improve target localization.

For affordance-based queries (e.g., "used for sitting", "used to cut"), IntentFuse achieves a success rate of 95% while LeRF only produces 83%, indicating enhanced grounding of functional semantics. The largest gain appears in negation-based queries (e.g., "not a bed"), where IntentFuse reaches 88% success compared to the baseline. Unlike LERF, which naively propagates the original query tokens into CLIP and often activates on negated objects, IntentFuse interprets user intent and explicitly excludes distractors during structured prompt construction.

To validate our method in real-world settings, we compute the F1-score as the harmonic mean of precision and recall for descriptive prompts. Our method achieves an overall of 11.8% relative improvement compared to LeRF. This highlights the ability of our method to generalize structured grounding to cluttered, real-world indoor conditions.

E. Limitations

Our work marks a first step toward integrating structured language understanding into 3D visual grounding without retraining. While IntentFuse enhances grounding through structured prompting, several constraints remain. The quality of grounding depends on the accuracy of LLM-generated semantic roles; ambiguous or under-specified queries can still lead to semantic drift and spurious activations. For example, a query like "something soft" may produce diffuse activations across multiple soft materials if the intended target (e.g., a pillow) is absent from the scene (Fig. 7). Although the context provider can mitigate such cases by supplying environment-specific priors (e.g., suggesting a pillow when present), its optional nature means user intent cannot always be fully disambiguated. Our evaluation is limited to LERF on the Replica dataset; future work will extend this analysis to stronger baselines (e.g., LangSplat, IGS2GS) and additional datasets. Similarly, exploring alternative LLMs (e.g., Gemini, LLaMA, Mistral) could improve robustness across backends. Finally, the current system does not yet support explicit spatial relationship modeling, limiting applicability to complex relational queries.

V. CONCLUSIONS

We introduced IntentFuse, a plug-and-play middleware that reformulates free-form language queries into structured, CLIP-aligned prompts for 3D representations such as LERF. Experiments on synthetic and real-world indoor scans show clear gains in grounding affordances, negations, and user intent over baseline LERF. Because IntentFuse operates atop any CLIP-based representation—whether volumetric, point-cloud, or image-based—it offers broad applicability. Future work will extend this approach to dynamic scenes, real-time robotics, and integration with scene reasoning modules to enable more natural human–agent collaboration.

REFERENCES

[1] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik, "Lerf: Language embedded radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19 729–19 739.

[2] E. Gomel and L. Wolf, "Diffusion-based attention warping for consistent 3d scene editing," *arXiv preprint arXiv:2412.07984*, 2024.

[3] M. Qin, W. Li, J. Zhou, H. Wang, and H. Pfister, "Langsplat: 3d language gaussian splatting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 051–20 060.

[4] W. Li, R. Zhou, J. Zhou, Y. Song, J. Herter, M. Qin, G. Huang, and H. Pfister, "4d langsplat: 4d language gaussian splatting via multimodal large language models," *arXiv preprint arXiv:2503.10437*, 2025.

[5] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[6] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM transactions on graphics (TOG)*, vol. 41, no. 4, pp. 1–15, 2022.

[7] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, "Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5855–5864.

[8] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[10] F. Liu, J. Liu, Z. Fang, R. Hong, and H. Lu, "Visual question answering with dense inter-and intra-modality interactions," *IEEE Transactions on Multimedia*, vol. 23, pp. 3518–3529, 2020.

[11] Q. Gu, A. Kuwajerwala, S. Morin, K. M. Jatavallabhula, B. Sen, A. Agarwal, C. Rivera, W. Paul, K. Ellis, R. Chellappa *et al.*, "Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 5021–5028.

[12] U.-H. Kim, J.-M. Park, T.-J. Song, and J.-H. Kim, "3-d scene graph: A sparse and semantic representation of physical environments for intelligent agents," *IEEE transactions on cybernetics*, vol. 50, no. 12, pp. 4921–4933, 2019.

[13] T.-T. Do, A. Nguyen, and I. Reid, "Affordancenet: An end-to-end deep learning approach for object affordance detection," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 5882–5889.

[14] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.

[15] P. Zech, S. Haller, S. R. Lakani, B. Ridge, E. Ugur, and J. Piater, "Computational models of affordance in robotics: a taxonomy and systematic classification," *Adaptive Behavior*, vol. 25, no. 5, pp. 235–271, 2017.

[16] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. Shamma, M. Bernstein, and L. Fei-Fei, "Image retrieval using scene graphs," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3668–3678.

[17] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5831–5840.

[18] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, M. Deitke, K. Ehsani, D. Gordon, Y. Zhu *et al.*, "Ai2-thor: An interactive 3d environment for visual ai," *arXiv preprint arXiv:1712.05474*, 2017.

[19] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik *et al.*, "Habitat: A platform for embodied ai research," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9339–9347.

[20] J. Straub, T. Whelan, L. Ma, Y. Chen, E. Wijmans, S. Green, J. J. Engel, R. Mur-Artal, C. Ren, S. Verma *et al.*, "The replica dataset: A digital replica of indoor spaces," *arXiv preprint arXiv:1906.05797*, 2019.

[21] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love *et al.*, "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024.

[22] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[23] X. Li, R. S. Cruz, M. Xi, H. Zhang, M. Perera, Z. Wang, A. Ravendran, B. Matthews, F. Xu, M. Adcock, D. Wang, and J. Liu, "Queryable 3d scene representation: A multi-modal framework for semantic reasoning and robotic task planning," 2025, accepted to ACM Multimedia 2025 (ACM MM), to appear.