MAGPIE: MEMORY-AUGMENTED AFFORDANCE GROUNDING VIA POINT-CLOUD INSTANCE SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Language-guided 3D affordance grounding requires a model to segment, directly in a 3D scene, the functional element that an instruction refers to. The leading methods answer it with scale: an 8B–72B vision–language model operates inside the grounding loop, every query incurs its forward-pass cost, and the decision rests on captured RGB-D frames in which the element may be occluded by its own carrier or absent from every view. Such models are designed to turn language into an executable goal, whereas the choice among instances of the same type depends on where each candidate lies relative to the others, a 3D signal that prior work largely leaves unused. MAGPIE therefore restricts the large models to a coarse region and settles the question in 3D. Specifically, a frozen vision–language model parses the instruction and selects a rendered view, a frozen grounder delimits the candidate set to one region of that view, and no large model is consulted again. MAGPIE stacks three lightweight heads on frozen point-cloud features: a type head proposes candidate instances, a set-selector resolves which same-category instance the instruction refers to, and a refine head tightens the selection into the final mask. The heads consume no image, captured or rendered, and disambiguate visually near-identical instances from 3D structure alone. On SceneFun3D, MAGPIE leads the strongest point-cloud-only prior work on every reported metric, surpasses peer-reviewed RGB pipelines up to 72B on AP_{50} and AP_{25}, and exceeds concurrent 30B systems on AP_{50} with three orders of magnitude fewer trainable parameters.

1 INTRODUCTION



(a) Where the instance decision is made.

(b) Breadth across three benchmarks.

Figure 1: **Overview.** (a) View-based methods route the decision through captured or rendered 2D views, in which a VLM selects the instance, whereas MAGPIE uses the VLM only for coarse region parsing and resolves the final choice among candidates in 3D. (b) MAGPIE against the strongest point-cloud-only baseline per benchmark: AffordBot-72B on AP, AP_{50}, AP_{25}, and mIoU of the SceneFun3D validation split, and PointRefer on the seen and unseen mIoU of full-view AGPIL and of LASO. Each axis is scaled independently and labeled with the raw scores of Tables 1, 6 and 7.

Scene-level affordance grounding requires a model to segment, in the 3D scan of a full indoor scene, the functional element that a natural-language instruction refers to, such as a handle, a knob, or a button (Delitzas et al., 2024). An embodied agent instructed to "open the middle left drawer of the dressing table" can act only after the sentence has been resolved into the points of one drawer handle. Within this setting, the real challenges begin with scale: across the 30 SceneFun3D validation scenes, target elements cover merely 0.19% of the scene points (Table A3), and any point-level predictor thus trains and predicts under a negative–positive imbalance of roughly $500{:}1$. The instruction offers little direct help: task descriptions are frequently ambiguous, in that they name neither the element nor an object category, and resolving them inherits the full spatial-reasoning burden of classical 3D grounding together with commonsense about how everyday objects are operated (Corsetti et al., 2025; He et al., 2025). The ambiguity sharpens where the scene repeats itself: nearly identical alternatives surround the target, selecting the wrong one fails the task outright, and the correct choice rests on

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

where each candidate sits among the others rather than on anything visible on the candidate (Wang et al., 2026a). Moreover, the target may not even be visible: an element of this size hides behind furniture or its carrier, and can be absent from every captured view. Nor does a vision–language model (VLM) offer an inexpensive remedy: image understanding at this granularity demands substantial capacity, and compact models fall short. A model must therefore find, among alternatives that appearance alone cannot distinguish, a target it can barely observe and the instruction never names.

Few existing lines of work confront these challenges jointly. Classical 3D visual grounding and open-vocabulary scene understanding operate at the granularity of whole objects (Chen et al., 2020; Takmaz et al., 2023; Huang et al., 2024; Jain et al., 2025; Koch et al., 2024; Gu et al., 2024), and the sub-percent, part-level target of an affordance instruction usually falls below their resolution. Object-level affordance learning presupposes a pre-localized shape and a class balance that scene-level input does not grant (Deng et al., 2021; Yang et al., 2023; Gao et al., 2025; Zhu et al., 2025; Mo et al., 2021). Nor does either family retain what distinguishes repeated elements: per-mask and per-view pooling flattens intra-object structure and the relative arrangement on which the choice among near-identical candidates rests. The recipes designed for the scene-level task route the decision through two-dimensional views: they lift per-view segmentations of captured frames into 3D (Corsetti et al., 2025; He et al., 2025; Wang et al., 2026a; Lin & Xu, 2026), or render 3D instance proposals for a multimodal model to select among (Wang et al., 2025), an approach bounded by the recall of the proposal generator it inherits. Under either route an element absent from every view remains unrecoverable, and errors committed per view propagate through the multi-stage pipelines. Across these scene-level pipelines, the semantic burden of the task largely rests with large VLMs, and the capacity that requires keeps inference costly. No existing route provides a mechanism by which the scene's own geometry, conditioned on the instruction, carries the decisive choice among candidates.

The mechanism missing from these pipelines is, however, largely present in the scene itself. Within a candidate region, frozen point-cloud features from a self-supervised encoder carry sufficient information to separate a functional element from its non-target surroundings, finely enough to delimit rather than locate it. Across scenes, moreover, elements that share an affordance tend to share geometric form, and affordance can thus be inferred from shape alone. The part-level perception on which current pipelines concentrate their capacity is therefore available from geometry without task-specific learning. The choice among near-identical siblings lies beyond what these features alone settle: within a repetitive arrangement they separate the sibling candidates only in part, and the decision falls instead to the instruction together with where each candidate stands among its neighbors. A linear probe on the frozen features recovers the 9-way affordance type at 0.846 accuracy (Table 4), and an oracle granted the correct cluster raises mIoU only from 24.9 to 26.6 on SceneFun3D (Table 2). Once perception is secured, what remains is a selection problem conditioned on language and relative arrangement, which demands far less capacity than the scene-scale image understanding through which the task is routed. Fig. 1 illustrates the contrast: the element the instruction designates is visually interchangeable with its siblings, and only its position within the local arrangement sets the target apart. Scale devoted to perception leaves that choice untouched, and resolving it requires a different form of learning, one that weighs the relative arrangement of the candidates against the instruction. A grounding model should therefore allocate its learned capacity to this choice.

Guided by this insight, we develop MAGPIE, a grounding framework that carries the full task, from proposing candidates to committing to one and delineating it, on the point cloud alone. The core of MAGPIE is a lightweight selector: it reads the candidates of a region as one set, weighs their relative arrangement against the instruction, and commits to a single instance. Fig. 2 isolates the gap it closes: a linear read of a candidate, scored independently of the others, picks the correct sibling on 50.9% of the multi-candidate queries, while our trained set-level selector succeeds on 88.6% (Section 4.4). The frozen features thus carry part of the sibling choice, and a lightweight read-out over the set and the instruction, rather than added scale, completes it. The rest remains frozen or lightweight. No captured image enters a trained module, no external proposal generator supplies the candidates, and no VLM takes part in the selection. A target absent from every captured view thus remains reachable, as the choice is made among 3D instances rather than within any view. On SceneFun3D, MAGPIE surpasses every point-cloud-only alternative on all reported metrics, with 47.2 AP_{25} against

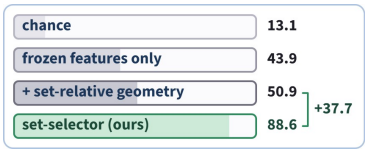

Figure 2: Sibling selection: linear probes on frozen features and set-relative geometry against the trained set-selector (see Table A9).

the 24.5 of the reproduced AffordBot-72B, and exceeds peer-reviewed RGB pipelines of up to 72B parameters on AP_{50} and AP_{25}. Selection is not yet saturated: an oracle choice over the same candidates still improves on our trained selector (Table 2), and headroom remains in candidate coverage and the choice within each region's candidate set. Our primary contributions are as follows:

- We design an affordance grounding pipeline that completes the full task, from proposals to the mask, on the point cloud alone. It removes the dependence on posed RGB-D imagery and external proposal generators, and with it their viewpoint and occlusion failures.
- We propose a set-level selector that resolves the choice among near-identical sibling instances from their relative arrangement and the instruction alone. This design places the learned capacity of MAGPIE at the single step where the difficulty of the task concentrates.
- We confine the VLM to upstream parsing and coarse region grounding and show that scene-level affordance grounding needs neither VLM scale nor captured imagery: lightweight heads over frozen geometry surpass peer-reviewed RGB pipelines on AP_{50} and AP_{25}.

2 Related Work

Since SceneFun3D (Delitzas et al., 2024) introduced scene-level affordance grounding, its pipelines have routed the decision through 2D views: a large VLM reads the instruction with captured frames (Corsetti et al., 2025; He et al., 2025; Wang et al., 2026a; Feng & Sabzevari, 2026; Lin & Xu, 2026) or with renders of finetuned Mask3D (Schult et al., 2023) proposals (Wang et al., 2025), and the part-level decision is lifted into the scan at the costs in capacity and visibility that MAGPIE avoids. General-purpose scene understanding pools one descriptor per mask, view, or graph node (Takmaz et al., 2023; Huang et al., 2024; Koch et al., 2024; Gu et al., 2024) and retains little of the arrangement among near-identical parts, whereas 3D referring (Chen et al., 2020; Achlioptas et al., 2020) models the candidates jointly through mutual attention under the instruction (Zhao et al., 2021; Chen et al., 2022; Jain et al., 2025), view rotation (Huang et al., 2022), or decoupled language (Wu et al., 2023), the relational form that the selector of MAGPIE carries to near-identical functional elements. Frozen features of the self-supervised Sonata (Wu et al., 2025) on Point Transformer V3 (Wu et al., 2024), strengthened by Concerto (Zhang et al., 2025), already recover strong semantic structure, as frozen features do in 2D (Oquab et al., 2024) and masked-latent pretraining does for language-conditioned 3D localization (Arnaud et al., 2025), and MAGPIE reads off these features the geometry-to-affordance mapping that object-level methods learn on isolated shapes (Deng et al., 2021; Yang et al., 2023; Gao et al., 2025; Zhu et al., 2025). Section J provides a more detailed survey.

3 Methodology

MAGPIE grounds a natural-language task description to the points of an interactive element in the scan of a complete indoor scene. Every judgment that determines the final mask takes place in 3D, carried by lightweight modules on frozen point-cloud features, while large pretrained models enter only where language must become an executable goal: they read the instruction and point at a coarse region in a rendered view of the point cloud. Fig. 3 overviews the framework, which comprises the following four stages: the **region-of-interest (ROI) selector** (Section 3.2), the **type head** with **feature-cut** proposals (Section 3.3), the **set-selector** (Section 3.4), and the **refine head** (Section 3.5).

3.1 Problem Formulation

MAGPIE takes as input the point cloud $\mathcal{P} = \{\mathbf{x}_i\}_{i=1}^{N} \subset \mathbb{R}^3$ of a complete indoor scan with N points, and a natural-language instruction q that names a task to carry out in the scene. The model returns a continuous score field $\mathbf{s} \in [0,1]^N$ over $\mathcal{P}$ whose support covers the functional element that must be physically operated to accomplish q, and only the evaluation protocol binarizes it. A frozen point-cloud encoder Φ (Zhang et al., 2025) maps the scan to per-voxel features $\mathbf{F} = \Phi(\mathcal{P})$ in one forward pass per scene, shared by every downstream module. Between stages R, G, S, and M of Fig. 3, z denotes the structured description of the instruction, π the camera of the selected view, B the candidate region in that view, $\mathcal{C}$ the set of K candidate instance clusters, and c^* the selected instance.

3.2 From instruction to candidate region

The ROI selector realizes stage R: it converts the instruction into the two products the geometric stages consume, a structured description and a spatial region that bounds the candidate search.

Query parsing. Query parsing normalizes the instruction into a structured description $z = (a, r, p, o, \sigma)$: the interaction type a among the nine of SceneFun3D (Delitzas et al., 2024), with overlapping types folded into one family, the root object r carrying the element, the operated part p, the ordinal o among same-type instances, and the spatial qualifier σ verbatim from the sentence. The

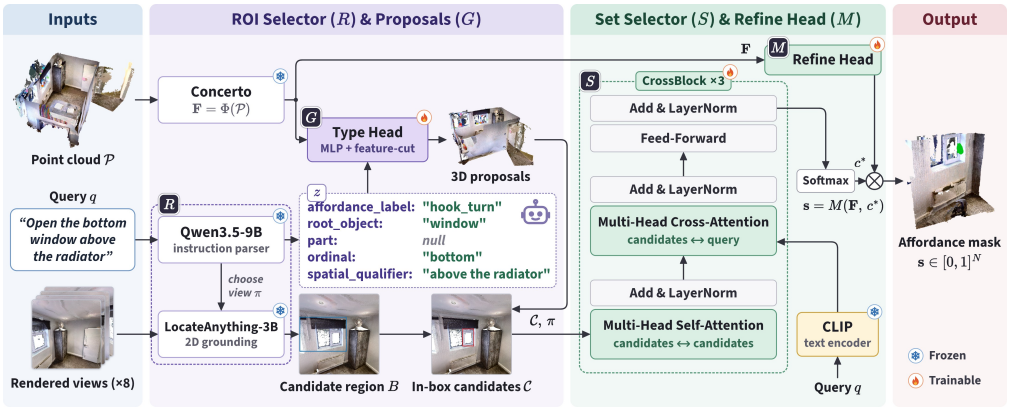

Figure 3: **MAGPIE overview.** The ROI selector (R) parses the instruction into a description z, a view π and a region B. The type head (G) proposes candidate instances $\mathcal{C}$, the set-selector (S) reasons over their geometry, relative arrangement and the language to commit to the referred instance c^*, and the refine head (M) turns it into the final mask $\mathbf{s}$. G, S and M train at probe scale on frozen models.

five fields form the parsed interface between the language and geometry ends of the pipeline, which confines linguistic ambiguity to a single stage (field-by-field spec in Sections A and D). Parsing responds to an entanglement: a single sentence carries the physical interaction, the object to locate, the fine element to operate, and the spatial context that separates repeated instances. A single grounding module would have to resolve all four in one step, although they rely on different cues.

View selection. View selection fixes the single rendered view, and with it the camera π, on which the region stage operates. MAGPIE renders a set Π of center-out RGB views of the scan and retains the one in which the frozen instruction parser judges the root object most visible, with no measurement on the scene, since interactive elements routinely hide behind furniture or their own carrier, and a poorly chosen view forces every later stage onto evidence wrong from the start. The camera π also fixes the only frame in which spatial qualifiers such as "left" and "right" acquire a definition.

Candidate region grounding. Region grounding delimits, on the selected view, the area where candidates are admitted. MAGPIE derives two complementary box families: root boxes localize the carrier from r and σ and deliberately circumscribe the whole piece of furniture, while target boxes localize the element from o, p, and σ, anchored to the root object. The candidate region takes their union $B = B_{\text{root}} \cup B_{\text{tgt}}$ on the selected view $\pi = R_{\text{view}}(\Pi, z)$. The *View* column of Table 3 measures how often the selected view exposes the referred element. The union favors containment over precision, since a target left outside B is unrecoverable while extra candidates are rejected downstream. Candidates are never detected in the image: B admits every 3D cluster of Section 3.3 whose centroid projects inside it, occluded or not. The frustum of B, however, imposes no bound along depth and thus admits clusters collinear with the target, which the selector screens out.

3.3 Affordance-type proposals on frozen features

Affordance-type proposals realize stage G: within the region that Section 3.2 delivers, they turn per-voxel geometry on the frozen features into the candidate set $\mathcal{C}$ handed to the selector. The type head rests on one observation: affordance types exhibit consistent geometric form across scenes. Elements that demand a hooked pull tend toward thin elongated bars, elements to be rotated toward round knobs, and the regularity holds largely independent of the furniture or scene carrying them. MAGPIE therefore learns, on the frozen features and without tuning them, a per-voxel classifier h_θ that maps geometry to affordance type. The head acts as memory that remembers what the requested interaction usually looks like, and it marks the voxels of the scene that look like it as follows:

$$h_\theta : \mathbf{f}_v \mapsto \mathcal{T} \cup \{\varnothing\}, \qquad \mathcal{V}_a = \{\, v : h_\theta(\mathbf{f}_v) \in \mathrm{fam}(a) \,\}, \tag{1}$$

where $\mathcal{T}$ denotes the interaction types, $\varnothing$ the background class, and $\mathrm{fam}(a)$ the family of the parsed type a. The background class absorbs the walls and floors dominating the voxel count: as an explicit rejection option, it spares the head from forcing every voxel into an interaction type.

Feature-cut turns the voxel set $\mathcal{V}_a$ into instances, the unit selection requires. Positional adjacency alone merges physically adjacent instances, two cabinet doors or a drawer column, into a blob. MAGPIE thus builds, over the candidate voxels, an adjacency graph demanding spatial proximity and feature similarity at once, and takes its connected components as the instance clusters as follows:

$$\mathcal{C} = \mathrm{CC}\big(\{\, (u, v) : \|\mathbf{p}_u - \mathbf{p}_v\| \le \rho,\ \mathrm{sim}(\mathbf{f}_u, \mathbf{f}_v) \ge \tau \,\}\big), \tag{2}$$

where ρ denotes the adjacency radius and τ the feature-similarity threshold, a cosine similarity between frozen features rather than a threshold on the score field. The construction rests on a measurable property: distinct physical instances, even in contact, exhibit a feature drop at their junction, and cutting edges by similarity converts that drop into a boundary. The threshold τ trades ease of selection against mask tightness, and Section 4.4 quantifies the trade across the sweep. The output $\mathcal{C}$, filtered by the region B of Section 3.2, is the candidate set that the selector receives.

3.4 SET-SELECTOR

The set-selector carries stage S and holds the core of MAGPIE: the choice among near-identical candidates of the same type, the failure source on which scene-level grounding concentrates. A candidate region typically covers a whole piece of furniture, hence the set $\mathcal{C}$ from Section 3.3 contains several sibling instances that look almost alike, and the task is to commit to the one the instruction refers to rather than to find the handles. The frozen features carry signal that separates siblings, which alone does not suffice. The criteria that decide a reference, qualifiers such as leftmost, lowest, or second, are properties of the candidate set. An isolated candidate cannot know its own rank, as that rank only exists once every member of the set is observed. Reading each candidate on its own therefore converts none of the available separability into a decision. The decision requires set context. MAGPIE accordingly abandons independent per-instance scoring and reasons over the set. Each candidate cluster becomes one token that concatenates four geometric quantities with one appearance quantity: the displacement of its centroid from the set centroid, the same displacement rotated into the selected view's frame, its height relative to the set mean in world coordinates, the logarithm of its size, and the mean frozen feature over its voxels. The tokens first self-attend, in which every candidate weighs its position against all the others, and then cross-attend to a token-level representation of the raw instruction q, which aligns the internal ordering of the set with the referring words. A scalar head then scores every candidate, a softmax over the candidate set trains under the cross-entropy of the annotated choice, and inference commits to the highest-scoring candidate as c^*. The prediction that results is relative by construction. The same candidate can receive a different score in a different set, which a qualifier such as leftmost demands and per-instance scoring cannot express. Each token encodes position in two frames, as left and right exist only relative to the observer whereas up and down do not, hence only the displacement is duplicated into the view frame. The language conditioning is token-level rather than pooled, which allows one candidate to align to the spatial qualifier while another aligns to the object noun. No large model takes part in this decision.

3.5 MASK REFINEMENT

The refine head closes stage M, from the selected instance to the final mask. Since selection has located the target, what remains is not coverage. Clusters formed by feature connectivity prove systematically looser than the annotated element, as they absorb the surfaces that carry it, the backplate of a handle or the base a knob is mounted on. The bottleneck at this stage is the boundary within an already located region, and the frozen features answer precisely that question. Within a local region they separate annotated from non-annotated voxels far more sharply than they separate one sibling from another, hence a tight mask is recoverable from the local geometry alone. The refine head decomposes mask prediction into two factors, tightness and localization. Tightness comes from a per-voxel foreground probability p_t, a classifier over the frozen features carrying one output channel per affordance type $t \in \mathcal{T}$, and the instruction selects the channel a that Section 3.2 parsed. Localization comes from the selection. The chosen cluster acts purely as a spatial gate, its axis-aligned bounding box dilated by a margin, and the final score multiplies the two and scatters back to the points as $s(\mathbf{x}) = \mathbb{1}[\,\mathbf{p}_{v(\mathbf{x})} \in \Gamma(c^*)\,] \cdot p_a(\mathbf{f}_{v(\mathbf{x})})$, where $v(\mathbf{x})$ denotes the voxel that contains point $\mathbf{x}$ and $\mathbf{p}_v$ the centroid of the scan points assigned to voxel v, $\Gamma(c^*)$ the dilated bounding box of the selected cluster, and $\mathbb{1}[\cdot]$ the indicator function. Since the per-type foreground probabilities p_t depend on neither the instruction nor the selection, a single refine-head pass per scene serves every instruction issued against it, and only the spatial gate changes between instructions. The refine head therefore never inherits selection errors in training, and its output channels are per type rather than shared, since the same geometry can afford a different interaction type in a different context.

3.6 TRAINING AND INFERENCE

Training updates only the three heads. The encoder and upstream models remain frozen, and each head is trained separately, on features extracted once in advance, under its own objective as follows:

$$\mathcal{L}_{\text{type}} = \text{CE}_{\text{cb}}\big(h_\theta(\mathbf{f}_v),\, y_v\big),\ \mathcal{L}_{\text{sel}} = -\log \frac{\exp(\ell_{k^\star})}{\sum_k \exp(\ell_k)},\ \mathcal{L}_{\text{ref}} = \sum_t \mu_{v,t}\, \text{BCE}\big(p_t(\mathbf{f}_v),\, m_{v,t}\big), \quad (3)$$

where y_v denotes the per-voxel type label over $\mathcal{T} \cup \{\varnothing\}$, $\mathrm{CE}_{\mathrm{cb}}$ a cross-entropy class-balanced against the background-dominated voxels (Section C), ℓ_k the selector scores with $k^\star$ the annotated instance, $m_{v,t}$ the mask label of voxel v on output t, BCE the binary cross-entropy, and $\mu_{v,t}$ a loss mask equal to one for a background voxel on every output and for an annotated voxel on its own type alone, hence background voxels are the only negatives. The stage interfaces define each head's target without reference to the others, hence the three optimizations neither interact nor require joint scheduling. Inference executes the same four-stage chain at fixed operating points, without test-time updates.

4 Experimental Results

4.1 Experimental Setup

Datasets. The primary benchmark is SceneFun3D (Delitzas et al., 2024), which annotates fine-grained functional elements in full indoor scans with element-level masks and natural-language task descriptions. We train on its 200 official training scenes and evaluate on all 445 instructions of the 30 validation scenes, as in prior work, since the annotations of the 85 test scenes are unreleased. Two object-level benchmarks, LASO (Li et al., 2024) and AGPIL (Zhu et al., 2025), probe whether the same architecture and frozen backbone transfer once localization and disambiguation vanish (Section 4.5). LASO spans 23 object categories and 18 affordance types, and AGPIL 17 types under full-, partial-, and rotation-view settings. All three benchmarks follow their official splits (Section B).

Metrics. Scene-level results follow the SceneFun3D evaluation protocol with seven measures: average precision (AP) averaged over intersection-over-union (IoU) thresholds from 0.50 to 0.95, its two single-threshold variants AP_{50} and AP_{25}, the matching recall measures AR, AR_{50}, and AR_{25}, and the mean IoU (mIoU). Scores are min–max normalized per query and binarized at the fixed mask threshold $\delta = 0.95$, and every measure is averaged over all queries. Object-level results adopt that literature's four standard measures, mIoU over twenty thresholds, the per-point area under the ROC curve (AUC), the histogram-intersection similarity (SIM), and the mean absolute error (MAE), computed with the official LASO implementation for an exact comparison to published numbers.

Baselines. The comparison covers the two input regimes that Table 1 separates. The sole prior point-cloud-only method is AffordBot (Wang et al., 2025), reproduced at 32B and 72B for the closest available comparison. The RGB-based regime contributes the representative scene-level pipelines Fun3DU (Corsetti et al., 2025), TASA (He et al., 2025), AffordMem (Wang et al., 2026a), and UniFunc3D (Lin & Xu, 2026) at 8B to 72B parameters, together with T-FunS3D (Feng & Sabzevari, 2026), which reads captured imagery without a VLM selection step. Object-level tables inherit the published rosters of each benchmark. The reproduced baselines follow their own protocols, which come closest to the published numbers. Both reproductions are insensitive to the operating point, and the lead of MAGPIE holds at every mask threshold (Section E): across the full range of δ, Fun3DU reaches at most 13.6 AP and 10.7 mIoU, against 22.7 AP and 22.3 mIoU for MAGPIE at $\delta = 0.95$.

Implementation. Every trained component reads only dense features from a frozen Concerto-large encoder (Zhang et al., 2025), a 207.7M-parameter Point Transformer V3 backbone, extracted once per scene and cached. The frozen upstream of Section 3.2 uses Qwen3.5-9B (Qwen Team, 2026) for parsing and view selection, LocateAnything-3B (Wang et al., 2026b) for the region boxes, and the CLIP ViT-B/32 text encoder (Radford et al., 2021) for the selector. The three heads hold 5.05M trainable parameters and train within five minutes on one RTX A6000. The efficiency claim thus concerns training and adaptation cost: a query at inference still consults the frozen 9B parser and 3B grounder upstream. Feature-cut uses $\rho = 5\,\mathrm{cm}$ and $\tau = 0.97$ at inference, and the gate of Section 3.5 dilates the selected box by 0.3 of each axis length plus 5 cm. Section C lists the hyperparameters, inference constants, and class-balancing scheme, and Section H the memory requirements.

4.2 Quantitative Results

Table 1 compares MAGPIE quantitatively against the state of the art on the SceneFun3D validation set and separates two input regimes: the upper half consumes the originally captured RGB-D frames, while the lower half operates on the point cloud alone. Within the point-cloud-only regime, MAGPIE leads on all seven metrics without captured imagery or a large VLM in the final instance selection. The closest comparison is the reproduced AffordBot-72B, the only prior method under the same input regime: MAGPIE improves AP by 7.5 points and AP_{25} by 22.7 points with a language model one-eighth the size. The AP_{25} margin indicates that the correct sibling is selected far more often, which is what the set-selector contributes, and the AP gain follows from the same selections. Against the RGB-based upper half, MAGPIE attains the highest AP_{50} in the table and exceeds the mIoU of TASA-72B and UniFunc3D-8B, while UniFunc3D-30B retains the lead on AP_{25}, mIoU, and relaxed-threshold recall, at the structural price that a pipeline reading captured frames cannot recover

Table 1: **Task-driven affordance grounding on the SceneFun3D validation split.** The upper block consumes captured RGB-D frames and the lower block operates on the point cloud alone. Bold marks the best entry per column within a block, and shaded rows are concurrent arXiv preprints. * marks results we reproduced, ‡ results we could not reproduce and therefore quote from the original paper. The last block gives MAGPIE the ground-truth affordance type and is excluded from the bolding.

Method	VLM sel.	RGB	AP	AP_{50}	AP_{25}	AR	AR_{50}	AR_{25}	mIoU
Fun3DU-9B	✓	✓	7.6	16.9	33.3	**27.4**	38.2	46.7	15.2
Fun3DU-9B*	✓	✓	5.6	11.0	21.6	27.4	35.7	44.0	10.7
TASA-72B‡	✓	✓	**23.2**	26.9	28.6	–	–	–	19.7
T-FunS3D	×	✓	8.1	17.8	34.5	23.8	35.8	46.9	15.7
AffordMem-32B	✓	✓	–	20.1	41.7	–	44.0	50.6	17.3
UniFunc3D-8B	✓	✓	–	23.8	44.0	–	46.1	55.5	20.9
UniFunc3D-30B	✓	✓	–	**31.2**	**51.0**	–	**47.0**	**58.9**	**24.3**
AffordBot-32B*	✓	×	8.8	12.1	17.3	9.0	13.3	17.3	8.9
AffordBot-72B	✓	×	15.5	20.0	23.3	–	–	–	14.0
AffordBot-72B*	✓	×	15.2	20.0	24.5	13.4	19.1	22.7	14.2
MAGPIE (ours)	×	×	**22.7**	**35.5**	**47.2**	**24.8**	**38.0**	**46.7**	**22.3**
(Oracle affordance type)									
MAGPIE (ours)	×	×	23.5	37.5	53.7	28.0	42.7	55.1	24.9

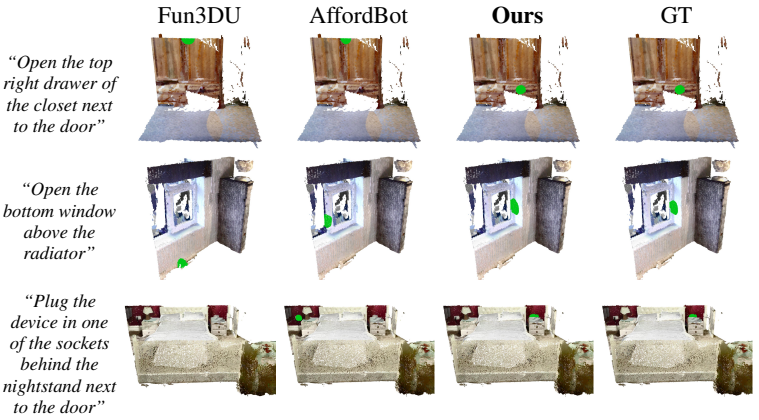

Figure 4: **Qualitative comparison on SceneFun3D.** Predicted affordance regions of each method in green on the reconstructed mesh, binarized at the mask threshold $\delta = 0.95$ used throughout. Each row shows one query whose target is identified by an ordinal or a spatial relation.

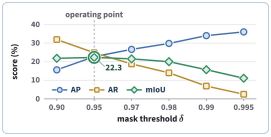

Figure 5: Mask-threshold sweep of Table A7: each measure as a function of δ, with $\delta = 0.95$ highlighted.

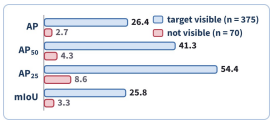

Figure 6: Grounding in Table A6 by target visibility in the selected view.

an instance no camera ever saw. On the recall side, masks tuned to sit tight around the ground truth trade coverage for precision, and AR lands within 2.6 points of Fun3DU. TASA-72B likewise holds a 0.5-point edge on the strictest AP, while MAGPIE leads it by 8.6 and 18.6 points on AP_{50} and AP_{25}. The ground-truth-type row replaces only the predicted label, and its 2.6-mIoU margin measures the cost of type-prediction error rather than a pipeline ceiling: it is borne entirely by the 174 mistyped queries, mostly confusions between *hook_pull* and *pinch_pull*, which carry 95% of the gap (Table A5).

4.3 Qualitative Results

Fig. 4 separates qualitatively the two capabilities behind the quantitative margins: the choice among same-type instances and the tightness of the final boundary. MAGPIE selects the specified instance in all three examples, at IoUs of 1.00, 0.62, and 0.30, the last with a mask that overshoots the outlet by 3.2×. AffordBot-72B identifies the correct affordance type and still places a plausibly sized mask on the wrong sibling in every example, the relative-reference ambiguity the set-selector resolves (Section 3.4). Fun3DU commits instead to the surrounding surfaces, the closet panel and the floor below the window, the mixture that feature-cut and the refine head separate (Sections 3.3 and 3.5). AffordBot's masks carry 82, 95, and 198 points against annotations of 81, 200, and 183, a failure of localization rather than of extent, while only 111, 93, and 2 points of Fun3DU survive the mask threshold. A stage-by-stage walk-through of five correct predictions and two examples of each failure mode appear in Section F, and Table 5 charges every validation query to the stage that fails it.

Table 2: **Ablation on the SceneFun3D validation split**, under the ground-truth type. *Sel.* denotes the instance-selection accuracy that Section 4.4 defines.

	Sel.	AP	AP_{50}	AP_{25}	AR	AR_{50}	AR_{25}	mIoU
Mask source ($\tau = 0.97$)								
MAGPIE (full)	69.3	**23.5**	**37.5**	**53.7**	**28.0**	**42.7**	**55.1**	**24.9**
w/o refine head	69.3	16.6	26.7	41.1	27.7	37.8	47.0	18.8
Feature-cut threshold τ								
off (xyz CC only)	**70.6**	20.8	32.8	48.1	28.0	41.1	52.8	23.1
$\tau = 0.95$	69.2	23.1	36.6	52.8	**28.5**	**43.4**	55.1	24.8
$\tau = 0.97$ (default)	69.3	23.5	37.5	53.7	28.0	42.7	55.1	24.9
$\tau = 0.99$	55.2	**25.2**	**39.3**	**54.4**	25.6	41.4	53.5	24.8
Selection ceiling								
box-oracle selection	–	24.9	40.0	57.5	30.7	47.4	59.3	26.6

Table 3: **Swapping the query-understanding VLM**, with the Concerto backbone and all downstream stages fixed. *Parse* measures affordance-type accuracy and *View* the fraction of the 445 queries whose selected view shows the target.

VLM	used by	Parse	View	AP	AP_{50}	AP_{25}	AR	AR_{50}	AR_{25}	mIoU
Qwen2.5-VL-32B	AffordBot-32B	0.598	0.802	21.2	32.6	43.2	**25.5**	36.6	43.8	21.3
Qwen2.5-VL-72B	AffordBot/TASA	0.611	0.805	22.3	34.2	44.5	**25.5**	37.3	46.1	21.8
Qwen3-VL-2B	–	**0.649**	0.647	16.9	26.5	35.3	19.4	29.0	35.7	17.4
Qwen3-VL-8B	UniFunc3D-8B	0.618	0.807	20.9	33.3	43.6	24.4	36.0	44.5	21.3
Qwen3-VL-30B-A3B	UniFunc3D-30B	0.605	0.760	19.9	31.2	41.4	23.0	33.7	42.3	19.8
Qwen3.5-2B	–	0.647	0.739	17.7	28.1	37.8	21.3	32.4	39.8	18.2
Qwen3.5-9B (ours)	–	0.609	**0.843**	**22.7**	**35.5**	**47.2**	24.8	**38.0**	**46.7**	**22.3**

Table 4: **Encoder ablation**, frozen Concerto-large against Sonata, under the predicted affordance type of Table 1. The probe columns read the frozen features before any head (Section G).

		SceneFun3D			Feature probe	
Method / encoder	Dim	AP_{50}	AP_{25}	mIoU	Margin	9-way
Concerto (Zhang et al., 2025) (ours)	1728	**35.5**	**47.2**	**22.3**	**0.129**	**0.846**
Sonata (Wu et al., 2025)	1232	27.9	40.9	16.1	0.076	0.767
AffordBot-72B (Wang et al., 2025)	–	20.0	23.3	14.0	–	–

Table 5: **Stage-wise outcome of the validation queries** under the ground-truth type. The rows partition the set, and the last two columns split each row by target visibility in the selected view.

Stage	Outcome	Queries	Share	Visible	Hidden
Type head + feature-cut	target absent from the candidate pool	58	13.0	39	19
ROI selector (frozen)	covering candidate outside the region	88	19.8	24	64
Set-selector	wrong candidate chosen	31	7.0	26	5
Correct instance recovered		**268**	**60.2**	231	37
Total		445	100.0	320	125

Table 6: **Language-guided affordance grounding on AGPIL, full-view setting.**

		Seen				Unseen			
Method	**2D**	**mIoU↑**	**AUC↑**	**SIM↑**	**MAE↓**	**mIoU↑**	**AUC↑**	**SIM↑**	**MAE↓**
3DAN	×	13.41	80.67	0.483	0.114	5.22	66.52	0.301	0.156
IAGNet	✓	20.51	84.85	0.545	0.098	7.95	71.84	0.352	0.127
OpenAD	×	19.93	85.84	0.587	0.094	7.81	73.75	0.370	0.119
PointRefer	×	20.02	87.66	0.595	0.091	8.13	75.45	0.385	0.118
LMAff3D	✓	21.23	**88.95**	**0.610**	**0.082**	**9.03**	**77.41**	**0.409**	**0.095**
MAGPIE (ours)	×	**22.70**	81.87	0.541	0.115	8.97	68.08	0.342	0.138

Table 7: **Language-guided affordance grounding on LASO.**

		Seen				Unseen			
Method	**2D**	**mIoU↑**	**AUC↑**	**SIM↑**	**MAE↓**	**mIoU↑**	**AUC↑**	**SIM↑**	**MAE↓**
ReferTrans	×	13.7	79.8	0.497	0.124	10.2	69.1	0.432	0.145
3D-SPS	×	11.4	76.2	0.433	0.138	7.9	68.8	0.402	0.158
ReLA	×	15.2	78.9	0.532	0.118	10.7	69.7	0.429	0.144
IAGNet	×	17.8	82.3	0.561	0.109	12.9	77.8	0.443	0.129
PointRefer	×	20.8	87.3	0.629	0.093	14.6	80.2	0.507	0.119
GEAL	✓	**22.0**	86.7	0.634	**0.092**	16.7	80.9	0.567	**0.106**
CMAT	✓	21.7	**88.2**	**0.681**	0.096	17.5	**82.9**	**0.625**	0.112
MAGPIE (ours)	×	21.4	82.0	0.561	0.124	**19.5**	80.9	0.529	0.125

4.4 Ablation Studies

The ablation studies trace the gains of MAGPIE to their sources, Tables 2 and 5 under the ground-truth affordance type and Tables 3 and 4 under the predicted affordance type of the main table.

Representation. The gains of MAGPIE do not hinge on a strong representation. Table 4 replaces the frozen 207.7M-parameter Concerto-large backbone with the 3D-only-pretrained 108.5M-parameter Sonata and retrains the three heads with an unchanged recipe, and MAGPIE still outperforms the strongest point-cloud-only baseline, AffordBot-72B, on every metric, with 27.9 against 20.0 AP_{50}, 40.9 against 23.3 AP_{25}, and 16.1 against 14.0 mIoU. The stronger representation raises these scores further: Concerto lifts mIoU to 22.3 and AP_{50} to 35.5, in step with the nine-way probe accuracy, from 0.767 to 0.846, and the part-versus-surroundings feature margin, from 0.076 to 0.129.

Refinement and selection. Refinement and selection contribute along separable axes: the refine head owns the boundary, the selector owns the choice, and the table dissociates the two. Its *Sel.* column counts, among the queries whose candidate pool contains a cluster covering the target (387 at the default τ; 391, 389 and 384 for off, 0.95 and 0.99), those whose selected cluster is the one of largest ground-truth overlap. Removing the refine head leaves the selection hit rate unchanged at 69.3 while mIoU drops by 6.1 points and AP_{25} by 12.6, hence coverage survives its removal and boundaries do not. Selection, in turn, still holds real headroom rather than sitting at a ceiling: an oracle choice over the same candidates reaches 26.6 mIoU against the 24.9 of MAGPIE, a gap of 1.7 mIoU points.

Candidate construction. Candidate construction decides the unit that selection operates on, and an easy-to-hit unit is not a well-segmented one. Removing feature-cut raises the selection hit rate to 70.6, the highest in the table, while mIoU falls to 23.1: merged neighbors form one blob that is easy to hit and loose to segment. The τ sweep places the operating point below the onset of fragmentation: mIoU stays at 24.8–24.9 for $\tau \in \{0.95, 0.97\}$, while $\tau = 0.99$ collapses the hit rate to 55.2 and, although AP rises to 25.2 on the smaller surviving masks, mIoU does not follow. The mask threshold

δ follows the same precision–coverage logic: AP rises and AR falls as δ tightens, and $\delta = 0.95$ attains the highest mIoU under both predicted and oracle types (Fig. 5 and Table A7).

Query understanding. Larger query-understanding VLMs improve view selection rather than parsing. Table 3 substitutes the query-understanding VLM with the downstream fixed: parse accuracy barely moves across a 36-fold parameter range, from 0.598 to 0.649, with the two 2B models parsing better than every larger one, while view selection rises from 0.647 to 0.843 and mIoU follows it from 17.4 to 22.3. Neither the 32B nor the 72B alternative surpasses the AP_{50} of the adopted 9B model. The 375 queries whose selected view shows the target reach 25.8 mIoU against 3.3 for the remaining 70 (Fig. 6 and Table A6), hence no stronger parser recovers a query whose view misses the target.

Stage-wise failures. Table 5 locates the failures stage by stage and assigns each of the 445 validation queries to the outcome that ends it, under the ground-truth type: a target is covered when a candidate cluster shares a voxel with it, and selected when the chosen cluster is the one of largest overlap. The bulk of the loss sits upstream of the trained choice: 13.0% of queries never receive a candidate cluster that covers the element, and a further 19.8% lose the covering candidate to the region. Only 7.0% reach the selector with the answer among the candidates and leave with the wrong one. Visibility is not a gate of its own, since the region is a frustum without an occlusion test, and the table counts a target as visible when at least 10% of its points pass a depth test against the scan in the selected view. Losing the region is likely for a hidden target, 64 of the 88 region failures, although not certain: 37 queries whose selected view does not show the target still end with the correct instance. Both larger modes lie in stages that hold no trained parameters of MAGPIE, which is where the remaining headroom concentrates: 146 of the 177 failures end before any candidate reaches the selector.

4.5 Object-Level Transfer: A Case Study

The object-level instantiation tests whether the architecture transfers unchanged once the task removes most of its stages: an already segmented object leaves no region to locate and no siblings to choose among, hence only the refine head is retained, with the same frozen backbone and the same partial supervision, and trained on the soft targets of this setting (Section C). Tables 6 and 7 evaluate this transfer with only the 1.02M parameters of the refine head trained and the affordance type predicted by the same 9B parser. The *2D* column and gray shading mark the methods that use 2D images or 2D foundation-model features besides the point cloud, and seen and unseen follow the official splits. Baselines are quoted from Zhu et al. (2025) on AGPIL and from Li et al. (2024) on LASO, with GEAL and CMAT from their own papers (Section E). Among methods that use no additional 2D imagery, the strongest of which is PointRefer (Li et al., 2024), MAGPIE leads on mIoU on all four reported splits, and every method whose mIoU exceeds that of MAGPIE relies on extra 2D images. The lead is specific to mIoU, as PointRefer remains ahead in ten of twelve AUC, SIM, and MAE entries. The pattern supports competitive scene-to-object transfer rather than a new object-level state of the art (Section E): on LASO, MAGPIE surpasses both GEAL (Lu et al., 2025) and CMAT (Huang et al., 2026) on unseen instances and trails them narrowly on seen ones, and on AGPIL full-view it leads LMAff3D (Zhu et al., 2025) on seen instances while nearly matching it on unseen ones. Table A9 isolates the choice the selector owns: scoring each candidate independently reaches 50.9% selection where the set-level selector reaches 88.6% (Section G). Section E extends the reading by affordance type, where lever handles reach twice the average mIoU, by view visibility, and to AGPIL under all three view settings, where partial views cost every method a fifth of its full-view mIoU.

5 Conclusion

This paper presented MAGPIE, a scene-level affordance grounding framework that carried the full task, from proposing candidate elements to committing to an instance and delineating its mask, on the point cloud alone. The design followed one observation: large pretrained models already supply strong perception, and the capacity a grounding model trains belongs at the choice among near-identical siblings that perception leaves open. MAGPIE placed that capacity in a lightweight set-selector over frozen point-cloud features; an affordance-type head proposed the candidate instances, and a refine head contracted the chosen one into the final mask. On SceneFun3D, MAGPIE led the point-cloud-only regime on all seven reported metrics and surpassed the reproduced AffordBot-72B by 7.5 AP and 22.7 AP_{25} with a language model one-eighth its size, while against RGB pipelines of up to 72B parameters it attained the highest AP_{50} overall. The same architecture transferred unchanged to object-level benchmarks, where the retained refine head remained competitive on LASO and AGPIL. The ablations traced each margin to its source: the selector owned the choice, the refine head owned the boundary, and the strength of the frozen backbone set the level of both.

AI USE STATEMENT

Generative AI assistants were used in three ways during this work. They were used as a sounding board while the design was taking shape, to discuss alternatives and to stress-test arguments; as a coding aid, for parts of the experiment and analysis scripts and for debugging; and for language editing of the manuscript. No experimental result, table or figure in this paper was produced by an AI assistant without being run, checked and interpreted by the authors, and no part of the scientific contribution was delegated. The authors take full responsibility for the content of this paper, including any text that was edited with such tools.

ETHICS STATEMENT

This work uses three public benchmarks, SceneFun3D, LASO and AGPIL, under their released licences, and introduces no new data collection. The scans contain indoor scenes of private homes; we use them exactly as distributed by their authors, add no annotations of our own, and do not attempt to identify people, places or possessions. No human subjects were involved and no user study was conducted. All pretrained components we build on are publicly released models, used frozen and without further training.

The intended application is an embodied agent that must act on a spoken or written instruction, and the foreseeable risk is therefore a mis-grounded prediction that causes a robot to operate the wrong element. Our own analysis is explicit that this happens: the method returns a confident mask on a sibling of the referred element in a measurable fraction of queries, and Fig. A1 quantifies where those failures arise. Any deployment should therefore treat the predicted mask as a proposal to be confirmed rather than as an executable command, and should not be used in settings where operating the wrong element carries a safety cost.

REPRODUCIBILITY STATEMENT

Anonymised code is available at `https://anonymous.4open.science/r/MAGPIE-0515/`. Section B states the datasets and splits, the exact frozen components, the order in which the stages are run and the evaluation protocol; Section C and Table A2 give the full training configuration of the three heads together with the inference constants; and Section H reports the hardware, the wall-clock and the memory the pipeline requires. The three heads train in five minutes on a single GPU from features that are extracted once, so the experiments can be repeated without the expense of the frozen encoder; the binding requirement is the memory of that single extraction pass, which needs a card of at least 48 GB at the voxel size we use.

REFERENCES

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. ReferIt3D: Neural listeners for fine-grained 3D object identification in real-world scenes. In *Computer Vision – ECCV 2020, 16th European Conference, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pp. 422–440. Springer, 2020. doi: 10.1007/978-3-030-58452-8_25.

Sergio Arnaud, Paul McVay, Ada Martin, Arjun Majumdar, Krishna Murthy Jatavallabhula, Phillip Thomas, Ruslan Partsey, Daniel Dugas, Abha Gejji, Alexander Sax, Vincent-Pierre Berges, Mikael Henaff, Ayush Jain, Ang Cao, Ishita Prasad, Mrinal Kalakrishnan, Michael Rabbat, Nicolas Ballas, Mido Assran, Oleksandr Maksymets, Aravind Rajeswaran, and Franziska Meier. Locate 3D: Real-world object localization via self-supervised learning in 3D, 2025.

Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. ScanRefer: 3D object localization in RGB-D scans using natural language. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

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 (NeurIPS)*, 2022.

Jaime Corsetti, Francesco Giuliari, Alice Fasoli, Davide Boscaini, and Fabio Poiesi. Functionality understanding and segmentation in 3D scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24550–24559, 2025. doi: 10.1109/CVPR52734.2025.02286.

Alexandros Delitzas, Ayça Takmaz, Federico Tombari, Robert Sumner, Marc Pollefeys, and Francis Engelmann. SceneFun3D: Fine-grained functionality and affordance understanding in 3D scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 3D AffordanceNet: A benchmark for visual object affordance understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Jingkun Feng and Reza Sabzevari. T-FunS3D: Task-driven hierarchical open-vocabulary 3D functionality segmentation, 2026.

Xianqiang Gao, Pingrui Zhang, Delin Qu, Dong Wang, Zhigang Wang, Yan Ding, and Bin Zhao. Learning 2D invariant affordance knowledge for 3D affordance grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3095–3103, 2025. doi: 10.1609/aaai.v39i3.32318.

Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. ConceptGraphs: Open-vocabulary 3D scene graphs for perception and planning. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

Lian He, Meng Liu, Qilang Ye, Yu Zhou, Xiang Deng, and Gangyi Ding. Task-aware 3D affordance segmentation via 2D guidance and geometric refinement, 2025.

Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3D visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15503–15512, 2022. doi: 10.1109/CVPR52688.2022.01508.

Yu Huang, Zelin Peng, Changsong Wen, Xiaokang Yang, and Wei Shen. Unlocking 3D affordance segmentation with 2D semantic knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Zhening Huang, Xiaoyang Wu, Xi Chen, Hengshuang Zhao, Lei Zhu, and Joan Lasenby. OpenIns3D: Snap and lookup for 3D open-vocabulary instance segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 169–185, 2024. doi: 10.1007/978-3-031-73033-7_10.

Ayush Jain, Alexander Swerdlow, Yuzhou Wang, Sergio Arnaud, Ada Martin, Alexander Sax, Franziska Meier, and Katerina Fragkiadaki. Unifying 2D and 3D vision-language understanding. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 26717–26739, 2025.

Sebastian Koch, Narunas Vaskevicius, Mirco Colosi, Pedro Hermosilla, and Timo Ropinski. Open3DSG: Open-vocabulary 3D scene graphs from point clouds with queryable objects and open-set relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Yicong Li, Na Zhao, Junbin Xiao, Chao Feng, Xiang Wang, and Tat-Seng Chua. LASO: Language-guided affordance segmentation on 3D object. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jiaying Lin and Dan Xu. UniFunc3D: Unified active spatial-temporal grounding for 3D functionality segmentation, 2026.

Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Dongyue Lu, Lingdong Kong, Tianxin Huang, and Gim Hee Lee. GEAL: Generalizable 3D affordance learning with cross-modal consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

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 (CVPR)*, 2022.

Kaichun Mo, Leonidas J. Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2Act: From pixels to actions for articulated 3D objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Toan Nguyen, Minh Nhat Vu, An Vuong, Dzung Nguyen, Thieu Vo, Ngan Le, and Anh Nguyen. Open-vocabulary affordance detection in 3D point clouds. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`. Used frozen at the 9B scale for query parsing and view selection.

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

Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3D: Mask transformer for 3D semantic instance segmentation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-vocabulary 3D instance segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Qirui Wang, Jingyi He, Yining Pan, Xulei Yang, and Shijie Li. Grounding by remembering: Cross-scene and in-scene memory for 3D functional affordances, 2026a.

Shihao Wang, Shilong Liu, Yuanguo Kuang, Xinyu Wei, Yangzhou Liu, Zhiqi Li, Yunze Man, Guo Chen, Andrew Tao, Guilin Liu, Jan Kautz, Lei Zhang, and Zhiding Yu. LocateAnything: Fast and high-quality vision-language grounding with parallel box decoding, 2026b. The 3B checkpoint is used frozen for region grounding.

Xinyi Wang, Xun Yang, Yanlong Xu, Yuchen Wu, Zhen Li, and Na Zhao. AffordBot: 3D fine-grained embodied reasoning via multimodal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer V3: Simpler, faster, stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, and Julian Straub. Sonata: Self-supervised learning of reliable point representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. EDA: Explicit text-decoupling and dense alignment for 3D visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19231–19242, 2023. doi: 10.1109/CVPR52729.2023.01843.

Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Grounding 3D object affordance from 2D interactions in images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Yujia Zhang, Xiaoyang Wu, Yixing Lao, Chengyao Wang, Zhuotao Tian, Naiyan Wang, and Hengshuang Zhao. Concerto: Joint 2D-3D self-supervised learning emerges spatial representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3DVG-Transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2908–2917, 2021. doi: 10.1109/ICCV48922.2021.00292.

He Zhu, Quyu Kong, Kechun Xu, Xunlong Xia, Bing Deng, Jieping Ye, Rong Xiong, and Yue Wang. Grounding 3D object affordance with language instructions, visual observations and interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17337–17346, 2025. doi: 10.1109/CVPR52734.2025.01616.