NEURAL INFORMATION PROCESSING SYSTEMS
2026 | POSTER

QGround: Condition-Wise Evidence Aggregation for 3D Grounding with 2D VLMs

Fengyun Wang[1] Jian Wang[2] Dingwei Zhang[3] Jinhui Tang[3] Qianru Sun[1]*

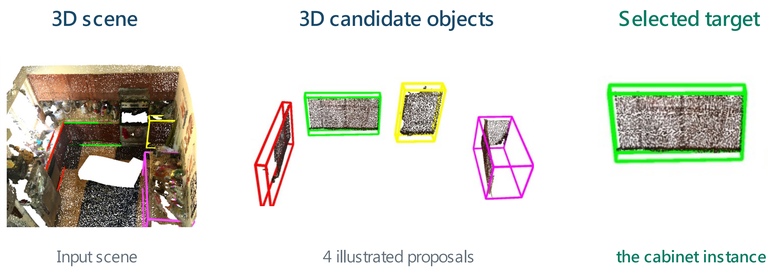
[1] Singapore Management University [2] Nanjing University of Posts and Telecommunications [3] Nanjing University of Science and Technology

CODE

3D Visual Grounding as Candidate Selection

Referring expression

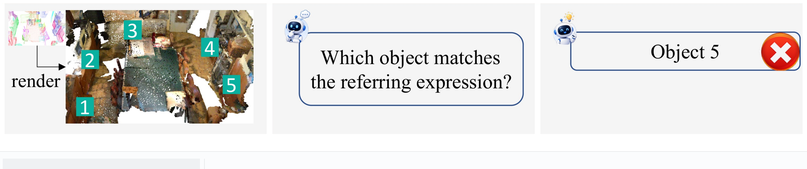
"the kitchen cabinets are left of the ovens; the kitchen cabinets are brown rectangles."



1 Why Holistic Grounding Decisions Fail

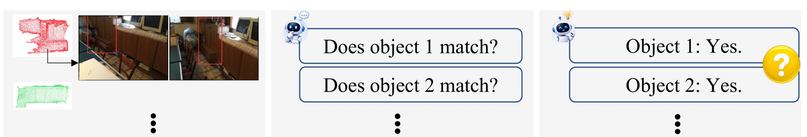
Referring expression: "the cabinet is below two screens; the cabinet is brown and rectangular."

1.1 Holistic Set Selection



How it works	Render a candidate set and ask the VLM to select one object in a single step.
Why it fails	One-shot set selection entangles visual grounding with ID selection over a rendered candidate set.

1.2 Holistic Candidate Matching

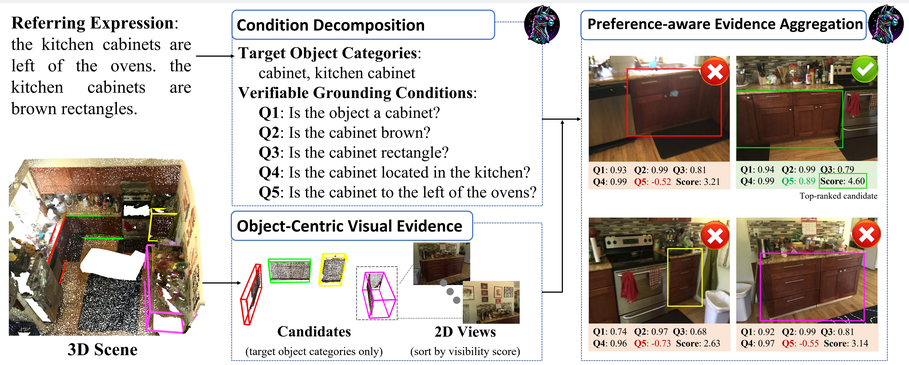


How it works	Keep candidates separate, but ask for one global match judgment per candidate.
Why it fails	Holistic matching obscures which category, attribute, and relational cues each candidate satisfies or violates, making partial matches difficult to disambiguate.

In both cases, the VLM's fine-grained judgments are not organized as comparable evidence across candidates.

2 Our Method - QGround

2.1 Overall Framework



2.2 Three Design Rules

Condition-Wise Verification

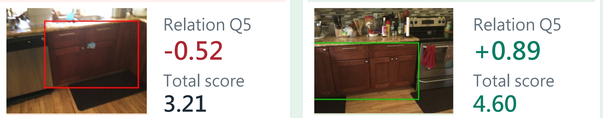
Isolate individual cues; do not compress the query into one overall match.

CATEGORY	ATTRIBUTE	RELATION
Q1: Is the object a cabinet?	Q2: Is the cabinet brown?	Q5: Is the cabinet to the left of the ovens?

Q3: rectangular shape · Q4: in the kitchen

Preference Reading

Read signed evidence score instead of hard YES/NO



The relation cue distinguishes two similar cabinet candidates.

Natural Inputs

Object-centric RGB evidence

Preserve the target and its surrounding scene context.

$$\mathrm{Vis}(o_i, v_k) = \frac{n_{\mathrm{visible}}}{n_{\mathrm{total}}}$$



Three natural scene views of the same marked candidate

2.3 Preference-Aware Evidence Aggregation

YES support	$p_{\mathrm{yes}}(o_i, q_j) = \sum_{v \in V_{\mathrm{yes}}} p(v \mid o_i, q_j)$
NO support	$p_{\mathrm{no}}(o_i, q_j) = \sum_{v \in V_{\mathrm{no}}} p(v \mid o_i, q_j)$
Signed evidence	$\mathrm{Score}(o_i, q_j) = p_{\mathrm{yes}}(o_i, q_j) - p_{\mathrm{no}}(o_i, q_j)$
Candidate selection	$\hat{o} = \arg\max_{o_i : c_i \in C} \sum_{j=1}^{M} \mathrm{Score}(o_i, q_j)$

3 Experimental Results

3.1 ScanRefer Val | Mask3D Proposals | Acc@0.5 (%)

Method	Supv.	Backbone	Unique	Multiple	Overall
3D-VisTA	FS	—	75.1	39.1	45.8
ConcreteNet	FS	—	82.1	38.4	46.5
SeeGround	TF	Qwen2-VL-72B	68.9	30.0	39.4
ZSVG3D	TF	GPT-4 turbo	58.4	24.6	32.7
VLM-Grounder	TF	GPT-4V	29.8	33.5	32.8
SeqVLM	TF	Doubao-1.5-pro	72.7	41.3	49.6
SPAZER	TF	GPT-4o	72.3	43.4	48.8
QGround	TF	Qwen2-VL-72B	72.0	41.4	47.3
QGround	TF	LLaVA-OV-7B	74.3	42.9	49.0
QGround	TF	LLaVA-OV-72B	77.6	46.0	52.2

FS = fully supervised | TF = training-free

Matched Qwen2-VL-72B: +7.9 pp vs SeeGround

3.2 Nr3D Val | Ground-Truth Candidates | Acc. (%)

Method	Easy	Hard	Dep.	Indep.	Overall
SeeGround	54.5	38.3	42.3	48.2	46.1
VLM-Grounder	55.2	39.5	45.8	49.4	48.0
SeqVLM	58.1	47.4	51.0	54.5	53.2
SPAZER	68.0	58.8	59.9	66.2	63.8
QGround 7B	69.0	49.6	50.7	63.3	59.1
QGround 72B	73.6	56.2	57.6	68.2	64.7

3.3 Ablation: Three Design Rules

Variant	Native	Cond.	Pref.	Unique @0.5	Multi. @0.5	Overall @0.5
Holistic baseline	✓	—	—	63.7	29.7	36.3
+ Condition-wise	✓	✓	—	72.0	37.9	44.5
+ Preference	✓	—	✓	69.7	38.1	44.2
QGround	✓	✓	✓	74.3	42.9	49.0
Rendered input	—	✓	✓	66.5	32.1	38.8

Full rule combination: +12.7 pp vs holistic; rendered input: □10.2 pp.

3.4 Comparison with Thinking-based VLM

Decision	Qwen3-VL Model	Unique @0.5	Multi. @0.5	Overall @0.5
Free-form thinking	Thinking-8B	64.9	34.2	40.1
QGround	Instruct-8B	77.2	45.0	51.2

Same Qwen3-VL family, distinct 8B variants; +11.1 pp overall.

3.5 Visual Comparison with SeeGround

