UZ3DVG: Unaided Zero-Shot 3D Visual Grounding with Generated Language Conditions

Supplementary Material

S1. Implementation Details

S1.1. Details of TSVDE

To complement high-level text-visual alignment features with fine-grained low-level visual details, thereby enhancing the representation of reasoning chain features and facilitating mapping to visual features, we propose the Text-Semantic-Guided Visual Detail Enhancement (TSVDE) module. At different alignment levels of the network, a progressive sampling strategy selects points from low-level, backbone-extracted point cloud features that exhibit high similarity to the text and incorporates them into high-level visual features. In contrast, at the highest layer where text-visual alignment has not yet been established, either purely random sampling or Farthest Point Sampling (FPS) is employed to preserve spatial coverage.

Let the low-level input consist of N points, indexed by $i \in \{1, \ldots, N\}$, where each point possesses 3D coordinates $\boldsymbol{x}_i \in \mathbb{R}^3$ and a visual feature vector $\boldsymbol{v}_i \in \mathbb{R}^D$. Stacking all visual features row-wise to obtain the matrix $\mathbf{V} \in \mathbb{R}^{N \times D}$, we first apply ℓ_2 normalization to both $\mathbf{V}$ and the text feature vector $\mathbf{T} \in \mathbb{R}^D$:

$$\tilde{\boldsymbol{v}}_i = \frac{\boldsymbol{v}_i}{\|\boldsymbol{v}_i\|_2 + \varepsilon}, \quad \tilde{\mathbf{T}} = \frac{\mathbf{T}}{\|\mathbf{T}\|_2 + \varepsilon}, \tag{S1}$$

where $\varepsilon > 0$ is a small constant for numerical stability. The cosine similarity is then computed as:

$$s_i = \tilde{\boldsymbol{v}}_i^\top \tilde{\mathbf{T}}, \tag{S2}$$

and a semantic relevance distribution is obtained via softmax:

$$\sigma_i = \frac{\exp(s_i)}{\sum_{j=1}^{N} \exp(s_j)}, \quad \sum_{i=1}^{N} \sigma_i = 1. \tag{S3}$$

Here, $\sigma_i \in [0, 1]$ quantifies the textual semantic relevance of point i.

We select the top $k_{\text{sem}}(l)$ points with the highest similarity scores, denoting their index set as $\mathcal{I}_{\text{sem}}$, and uniformly sample $k_{\text{rand}}(l)$ additional points from the remaining pool, with index set $\mathcal{I}_{\text{rand}}$. To balance semantic fidelity and spatial diversity across different alignment levels, we define level-dependent sampling proportions: $\big(\rho_{\text{sem}}(l),\, \rho_{\text{rand}}(l)\big), \quad \rho_{\text{sem}}(l) + \rho_{\text{rand}}(l) = 1$, where $\rho_{\text{sem}}(l)$ increases and $\rho_{\text{rand}}(l)$ decreases as the alignment becomes stronger. Here, ρ_{sem} and ρ_{rand} denote the semantic and random sampling proportions, respectively; l indexes the alignment level; and the random branch uses uniform or FPS sampling.

Given a total sample size K (with $K < N$), the numbers of semantically and randomly sampled points satisfy:

$$k_{\text{sem}}(l) = \lfloor \rho_{\text{sem}}(l) \cdot K \rfloor, \quad k_{\text{rand}}(l) \quad = K - k_{\text{sem}}(l). \tag{S4}$$

This strategy preserves spatial coverage in unaligned layers, introduces weak semantic guidance during partial alignment, and balances semantics and spatial diversity in fully aligned stages.

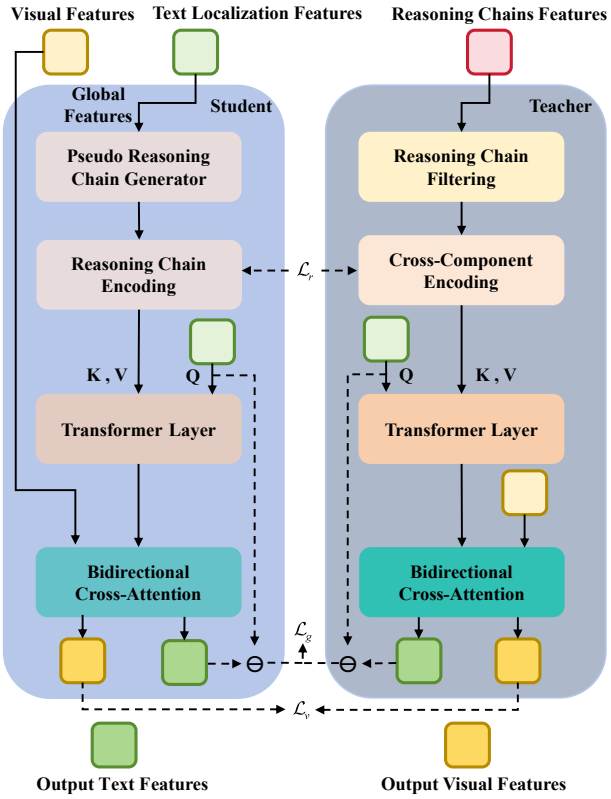

Figure S1. Overview of the distillation framework, where the student network learns from the teacher's reasoning chain.

Consequently, visual regions highly similar to the textual description receive richer detail representation, thereby more effectively supporting reasoning chain feature representation.

S1.2. Details of Reasoning Chain Distillation

To effectively transfer reasoning chain knowledge from the teacher network to the student, we design a reasoning chain distillation module, as illustrated in Figure S1. The teacher extracts reasoning logic and structural information from reasoning chains. The student approximates the teacher by generating three pseudo reasoning components from the spatial description features, with the goal of extracting reasoning-related knowledge from textual descriptions to support reasoning and localization.

As shown in Figure S1, after extracting the reasoning chain components, the teacher network applies a quality gating module to suppress unreliable components, then models component interactions with a self-attention layer and further grounds these components to the textual features via cross-attention. These processed teacher reasoning components serve two purposes: (1) they

provide component-level supervision for guiding the three pseudo reasoning components in the student, and (2) they are injected into the textual features to reinforce reasoning-relevant textual representations. The resulting reasoning-enhanced textual features are subsequently aligned with visual features via a bidirectional cross-attention module, producing paired textual and visual representations that link reasoning chains to visual content.

The student network first extracts a global context from the textual grounding description and uses it to initialize a pseudo reasoning generator composed of an MLP and an attention extractor. This generator produces three pseudo reasoning components, which are further aggregated into a global student reasoning representation. During distillation, the three pseudo reasoning components are aligned with the three teacher reasoning components, while the aggregated student reasoning representation is aligned with the corresponding global teacher reasoning representation.

To achieve high reasoning efficiency, we adopt a lightweight pathway analogous to that of the teacher. Specifically, the student first uses the pseudo reasoning components to enhance the textual grounding features through attention-based reasoning injection, and then further fuses the aggregated global reasoning representation into the text features. The resulting reasoning-enhanced textual features are subsequently aligned with visual features through a bidirectional cross-attention module.

To further facilitate the student's imitation of the teacher, we perform distillation at both the reasoning and feature levels. Specifically, the teacher's reasoning components supervise the student's pseudo reasoning components, while the teacher's reasoning-enhanced textual features and aligned visual features guide the corresponding student features through textual and visual feature distillation. This supervision enables the student to learn teacher-like reasoning features from the grounding description, internalize through attention the ability to extract reasoning-related information from the text, form reasoning-enhanced textual representations, and establish cross-modal correspondences between reasoning and visual content. Consequently, during inference, the student can generate pseudo reasoning components solely from the original grounding description to facilitate reasoning, text understanding, and text-vision alignment, thereby improving spatial reasoning capability.

To further ensure the quality of reasoning chain distillation, the training process is divided into three stages: warm-up, distillation, and fine-tuning. The teacher network follows a warm-up-then-distillation strategy. During the warm-up stage, the student network is optimized only under the supervision of pseudo labels, whereas the teacher network learns to extract knowledge from the reasoning chain, enhance the text grounding description, and subsequently align it with visual features. In terms of optimization, the teacher network is mainly driven by the detection loss and the localization loss. Specifically, the teacher loss consists of a vote loss for object center localization, an IoU loss for bounding box regression refinement, a classification loss for category discrimination, and an alignment loss for strengthening text-visual correspondence. Together, these terms constitute the overall teacher loss. After the warm-up stage, training proceeds to the distillation and fine-tuning stages. During these stages, the teacher continues to update itself using the same loss, while the student, in addition to optimization with the main task objective, is further aligned with the teacher through visual feature distillation, text enhancement distillation, and reasoning chain component feature distillation. As a result, the student learns the teacher's reasoning enhanced text-visual mapping and alignment patterns, thereby improving its ability to model reasoning information in grounding descriptions and ultimately enhancing grounding performance.

S1.2.1. Details of MSAG-RC

To generate 3D object-spatial relation description pairs and their reasoning chains for referential grounding in 3DVG, we propose the Open-Vocabulary Multi-Source Spatial Annotation and Reasoning Chain Generator (MSAG-RC). Our method takes video frame sequences from ScanNet scenes as raw input and samples one frame every 30 frames for object detection. For each sampled frame, spatial descriptions are automatically generated from the detected objects, forming structured description-reasoning pairs that support open-vocabulary 3DVG.

Given a scene image frame I_i $(i = 1, 2, 3, \dots)$, we use Grounding DINO [20] to obtain candidate object categories and 2D bounding boxes. The pipeline first filters these boxes by size, aspect ratio, and invalid categories, and then refines the remaining regions with SAM2 [29]. The refined masks are lifted to 3D using depth maps and camera parameters, from which axis-aligned 3D boxes are computed. These pseudo 3D candidates are then filtered by a quality evaluator for subsequent description generation.

To improve the vision-language model (VLM)'s [12] understanding of spatial relationships and enhance description quality, we convert the pseudo 3D box information into a structured prompt template, as shown in Table S1.

Let the pseudo-3D center of the i-th object be $\mathbf{c}_i = (x_i, y_i, z_i)^\top$, and its size be $\mathbf{s}_i = (s_{x,i}, s_{y,i}, s_{z,i})^\top$. Its volume is computed as:

$$V_i = s_{x,i} \cdot s_{y,i} \cdot s_{z,i} \tag{S5}$$

Based on V_i, we define the size hint as:

$$\text{size\_hint}_i = \begin{cases} \text{“large”}, & V_i > 1.0 \\ \text{“small”}, & V_i < 0.1 \\ \text{“medium”}, & \text{otherwise} \end{cases} \tag{S6}$$

The position hint is formatted as:

$$\text{pos\_hint}_i = \text{“}x = x_i\text{m},\ y = y_i\text{m},\ z = z_i\text{m”} \tag{S7}$$

These hints are embedded into the prompt template to provide contextual scene information, enabling the VLM to reason about spatial relationships between objects using their pseudo 3D centers and scales. To further guide the model in understanding inter-object spatial relations and generating high-quality spatial descriptions, we assign unique identifiers and annotate category labels directly on the 2D detections. The annotated image, together with the prompt template and pseudo 3D spatial context, is then provided to the VLM. Guided in this manner, the VLM generates human-like referential descriptions for object retrieval, producing natural language expressions that encode spatial relations and reasoning-based localization cues. Examples are shown in Table S1.

Diversity Control. To enhance the diversity of descriptions for the same object, we perform diversity augmentation on high-confidence 2D detections. Specifically, we generate multiple de-

Table S1. Prompt template for generating spatial relation descriptions and reasoning chains of 3D objects to guide referential grounding in 3DVG.

You are an expert 3D Scene Captioner with a built-in Self-Correction Verification module. For `obj{object_id}` (`{category}`), provide:

Given Information:

`Target:`	`obj{object_id}` (`{category}`)
`3D Position:`	{`pos_hint`}
`Size:`	{`size_hint`} ({s_x:.1f} × {s_y:.1f} × {s_z:.1f} m)
`Scene:`	{`scene_context`}
`2D image marked:`	{`image_url`}
`Variations:`	Use different sentence structures; focus on distinct spatial relations;...

Required Output:

1. **DESCRIPTION:** (15–40 words)
   Natural location description using room-relative terms. Include visual attributes.
2. **REASONING CHAIN:**
   After generating the description, provide a reasoning chain with structured localization logic:

`ANCHOR:`	The target object phrase itself
`REFERENCES:`	2–4 core localization phrases extracted from DESCRIPTION and explicitly bound to ANCHOR
`REASONING:`	One concise localization-logic statement derived only from DESCRIPTION
`VERIFICATION:`	Final self-verification that confirms the target and rejects distractors
`CONFIDENCE:`	High/Medium/Low

Format:

```
DESCRIPTION: [your description here]
```

REASONING CHAIN:

`ANCHOR:`	[target object phrase]
`REFERENCES:`	[[ANCHOR] relation phrase 1; [ANCHOR] relation phrase 2; ...]
`REASONING:`	[[ANCHOR] is the candidate satisfying the reference constraints; candidates violating any relation are distractors.]
`VERIFICATION:`	[final confirmation and distractor rejection]
`CONFIDENCE:`	[level]

Example:

DESCRIPTION:
The tv stand is against the wall, directly below a bulletin board covered in papers, featuring multiple white drawers.

REASONING CHAIN:

`ANCHOR:`	Tv stand.
`REFERENCES:`	[Tv stand] against the wall; [tv stand] directly below a bulletin board covered in papers.
`REASONING:`	[Tv stand] is the candidate that is against the wall and directly below the bulletin board covered in papers; candidates violating either relation are distractors.
`VERIFICATION:`	Confirm the large white stand with a brown top and multiple drawers; exclude other cabinets that do not satisfy both spatial relations.
`CONFIDENCE:`	High

scriptions by dynamically introducing variation cues, such as altering sentence structure, using synonyms for descriptors, emphasizing different visual attributes, focusing on distinct spatial relations, varying narrative tone, and selecting alternative reference

points. This strategy emulates the natural variability in human descriptions of object locations during referential grounding, thereby improving both the quality and diversity of the generated localization descriptions.

Reasoning Chain Construction. After obtaining the object localization descriptions, we prompt the VLM to simulate how a robot would locate the referred object in the current visual scene using the description, while structuring its reasoning process into three key elements, namely Anchor, Reference, and Reasoning, to provide more stable guidance for student model distillation. The `ANCHOR` must be the target object phrase itself, `REFERENCES` must contain 2–4 core localization phrases explicitly anchored to that target, and `REASONING` must be a single concise localization statement derived from those phrases. In addition, the output includes `VERIFICATION` for distractor rejection and `CONFIDENCE` in `High`/`Medium`/`Low`.

To further improve the quality of both the description and the reasoning chain, we use prompt engineering to instruct the VLM to perform reverse verification on both. A strict verifier independently tests whether the `DESCRIPTION` alone and the `REASONING_CHAIN` alone can uniquely localize the target object. If either test fails, the sample is regenerated up to a fixed number of attempts, and only samples passing both tests are retained; otherwise, they are discarded. If the target is successfully matched, the chain is assigned a confidence label of "`high`"; otherwise, the VLM regenerates the chain until a match occurs or the maximum number of attempts is reached. Unverifiable chains are labeled "`low`". These confidence scores guide the weighting of reasoning chains during model training, enabling smaller models to more effectively acquire the VLM's spatial reasoning logic from textual representations.

S1.3. Details of Mask3D Usage

In this experiment, our method UZ3DVG can be trained and evaluated entirely without relying on Mask3D [30]. The original results of our model on the ScanRefer dataset (i.e., Non-Mask3D) are reported in Table 1. However, for a fair comparison with prior methods such as ZSVG3D [51] and SeeGround [17], which use Mask3D-derived segmentation masks to refine localization boundaries, we adopt the same Mask3D usage protocol as SeeGround [17]. Specifically, on the ScanRefer dataset, we select the Mask3D segment corresponding to the region predicted by UZ3DVG's initial localization output and use it as the final boundary-refined prediction. The results after Mask3D-based refinement are also reported in Table 1.

S2. Additional Visualization Results

To further investigate the localization performance of our reasoning chain distilled model relative to the baseline method in descriptions containing complex reasoning logic, we conducted visual comparison experiments on localization descriptions with intricate spatial reasoning.

As shown in rows (a) and (e) of Figure S2, the description is a typical example involving multiple spatial reference points and complex positional relations. When the input description contains complex spatial positional relations, the baseline method SeeGround [17] consistently localizes the wrong target, whereas our approach accurately identifies the intended object. From rows (b)–(d), we observe that even when the description involves multiple directional references, our method still effectively localizes the target object, while the baseline fails to do so. This demonstrates the robustness of UZ3DVG in complex scenarios. This advantage stems from the reasoning chain distillation network, which enables the student network to mimic the teacher network in learning reasoning logic, thereby allowing it to extract and construct reasoning chains from localization descriptions to assist in object localization. Meanwhile, the Geometry-aware Spatial Modeling module (GeoSM) provides comprehensive global and local geometric information, facilitating enhanced mapping between the reasoning chain augmented text and visual spatial relationships, thereby improving localization performance.

Figure S2. Qualitative comparison between our method UZ3DVG and the baseline SeeGround [17] on referring expressions involving complex reasoning logic.