GAUGE3D: COORDINATE-GAUGE CONSISTENCY FOR DENSE 3D VISION–LANGUAGE GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Geometry-aware vision–language models ground language in 3D scenes, yet high accuracy in one coordinate frame can conceal sensitivity to arbitrary coordinate choices. Changing the global origin or orientation preserves the observed scene but requires dense masks, grounded text spans, and language answers to follow consistent transformation rules. We introduce GAUGE3D, a framework for evaluating and learning coordinate consistency in dense 3D vision–language grounding. The key idea is to turn coordinate re-expression into paired supervision of the complete output interface. A geometry-only compiler transforms camera poses, world-space coordinates, and geometric targets while preserving the underlying color and depth observations. A joint objective constrains mask equivariance on a common source support, text-span grounding consistency, and answer invariance for frame-independent queries, with transformed targets for frame-dependent queries. A consistency audit activates paired training when repeatable discrepancies exceed the identity and compiler noise floor, linking diagnosis to targeted supervision without adding a new 3D encoder. Evaluations on ScanRefer, ScanQA, and SQA3D show that GAUGE3D improves both dense grounding and spatial question answering over the underlying 3D vision–language model.

1 INTRODUCTION

Understanding and reasoning about 3D environments is fundamental for embodied AI and spatially grounded applications. Grounded 3D language tasks have established natural-language object localization and fine-grained object identification as core settings for reasoning over real-world scans (Chen et al., 2020; Achlioptas et al., 2020). Recent 3D vision–language models (3D-VLMs) adapt 2D multimodal backbones to posed RGB–D streams (Lin et al., 2026). A released world-space model lifts features into world coordinates, pools nearby observations, adds four-component 3D rotary positions, and feeds a mask decoder (Lin et al., 2026). This geometric interface connects language to spatial support, making the relationship between coordinate representations and physical objects central to how its outputs should be interpreted.

Yet, a world coordinate is a representation choice, not a physical observable. The same RGB–D observations can be expressed under many global origins and axis conventions. Equivariant 3D architectures have shown how geometric representations can respond predictably to roto-translations (Fuchs et al., 2020), but this principle must also hold at the grounded language interface, where masks, text spans, and answers have different transformation rules. Evaluating a model that uses world coordinates therefore requires both grounding accuracy in a given frame and predictable output transformations across coordinate gauges. This raises a basic question: does the complete output interface transform predictably under coordinate re-expression, or can a high grounding score coexist with coordinate-dependent masks, spans, or answers?

Figure 1 illustrates why accuracy in one coordinate frame is insufficient: with RGB–D observations fixed, coordinate re-expression should preserve the object identified by inverse-mapped masks and grounded text, while answers about object identity should remain unchanged and world-space positions transform with the coordinates. These requirements link what the model segments, which object its words refer to, and what it answers. Consistency must therefore be evaluated across the complete interface.

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.

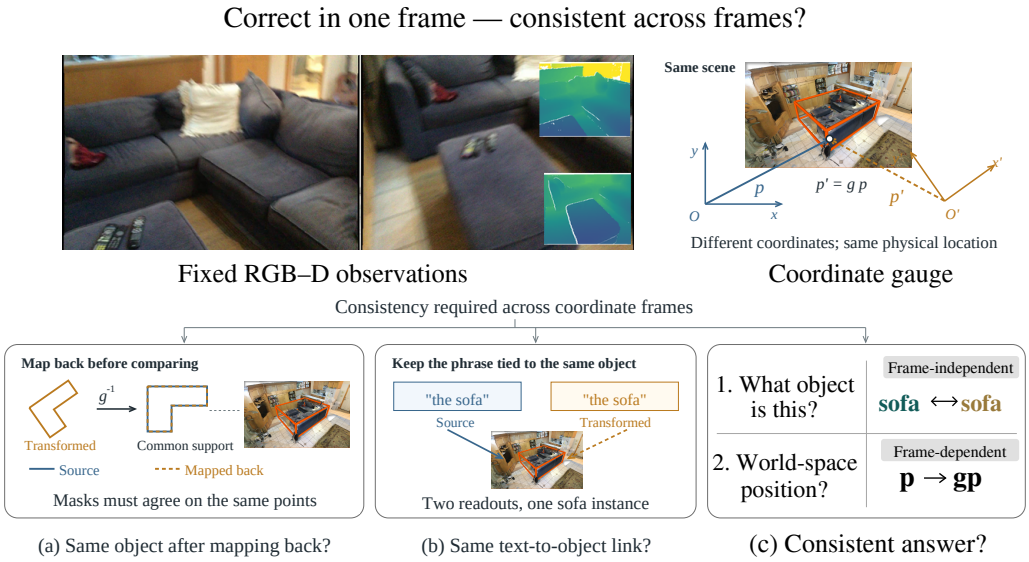

Figure 1: **Coordinate-gauge consistency for fixed RGB–D observations.**

Recent evaluation studies ask whether 3D models rely on evidence rather than language shortcuts and whether annotations remain answerable under the available frame (Ma et al., 2026; Zhang et al., 2026a). They make evidence dependence visible, but leave coordinate re-expression outside the evaluation of a dense world-space interface. Holding the observations fixed isolates this question from changes in visibility or available evidence. A valid coordinate pair supplies both the same physical scene and an explicit rule for relating its outputs. Ordinary rotation augmentation alone does not establish this equivariance property. The paired outputs must also be compared under that rule.

We propose GAUGE3D to make this coordinate contract executable through paired examples that differ only by a valid global coordinate transformation. The model maps posed observations and a query to a dense mask, a grounding score, and a language answer. A geometry-only compiler jointly transforms camera poses, scene coordinates, and geometric targets while preserving the RGB–D observations. It verifies the projection and inverse-mapping relationships before the pair reaches the model, establishing the correspondence needed to compare outputs on the same source points. For frame-dependent queries, the semantic target is transformed with the same declared rule.

The same correspondence connects diagnosis to supervision. The audit measures mask agreement after mapping back, text association with matched instances, and answer consistency under the query-specific rule. The proposed paired objective retains the original supervised losses and is activated by the baseline audit's repeatable excess gap, linking the consistency contract to a targeted intervention. Its three terms supervise these same output relationships without adding a new inference-time encoder. Ordinary grounding and answer accuracy remain separate measures, keeping coordinate consistency tied to task utility.

Our contributions are as follows:

- We formulate coordinate-gauge consistency at the complete output interface of a released world-space 3D vision–language model, separating invariant language outputs from equivariant geometric outputs.
- We propose a geometry-only paired compiler that transforms poses, world coordinates, and mask targets under an explicit group action while preserving the underlying RGB–D observation, together with an audit-gated joint evaluation of mask-cycle, text-span, and answer consistency.
- We validate GAUGE3D on ScanRefer, ScanQA, and SQA3D, demonstrating the effectiveness of the proposed framework across these benchmarks.

2 Related Work

3D language grounding and world-space models. ScanRefer and ReferIt3D established referring-expression localization and fine-grained instance identification in RGB–D scans (Chen et al., 2020; Achlioptas et al., 2020). Later 3D vision–language systems aligned point-cloud representations with language for grounding and question answering (Zhu et al., 2023; 2024). 3D-LLM and LLaVA-3D brought 3D features into pretrained language backbones; Qwen-3D adds world-space tokens and dense mask decoding (Hong et al., 2023; Zhu et al., 2025; Lin et al., 2026). ScanQA and SQA3D extend these settings to scene-level question answering (Azuma et al., 2022; Ma et al., 2023). These works mainly assess prediction accuracy in the supplied frame. GAUGE3D tests whether the same referent, mask, text span, and query-conditioned answer follow their required transformations under an equivalent coordinate description.

Spatial consistency and coordinate equivariance. Spatial-consistency benchmarks extend evaluation beyond static views: SCOPE tests allocentric and egocentric reasoning across viewpoints; SpinBench probes perspective and rotation; Real-3DQA and ReVSI examine answerability and evidence validity (Kim et al., 2026; Zhang et al., 2026b; Ma et al., 2026; Zhang et al., 2026a). These studies vary viewpoint or evidence rather than only the coordinate chart of fixed RGB–D observations; camera motion changes visibility, whereas global re-expression does not. CoordRefer conditions 3D box regression on a selected reference frame (Li et al., 2026), but does not target full-output cycle consistency across equivalent global frames. SE(3)-Transformers model roto-translation equivariance in geometric features (Fuchs et al., 2020). Recent work studies gauge-aware rotary attention and projective or reprojection consistency in view synthesis and 3D vision–language features (Li, 2026; Wu et al., 2026; Hariat et al., 2026). These approaches do not define joint transformation rules for masks, text-span associations, and answers. GAUGE3D preserves RGB–D evidence, transforms poses and targets by an explicit group action, and evaluates mask-cycle equivariance, span grounding, and query-aware answer consistency under invariant and frame-dependent rules.

3 GAUGE3D

3.1 Input and output contract

Let an example contain posed RGB–D observations

$$X = \{(I_v, D_v, K_v, T_v)\}_{v=1}^{V},$$

where I_v and D_v are RGB and depth images, K_v is the camera intrinsic matrix, and T_v maps camera coordinates to a scene coordinate frame. A query q can request an object, relation, count, or metric quantity; a dense grounding model produces

$$\hat{Y} = F_\theta(X, q) = (\hat{M}, \hat{s}, \hat{a}),$$

where $\hat{M}$, $\hat{s}$, and $\hat{a}$ are mask, span-score, and answer outputs.

The gauge audit separates $\mathcal{Q}_{\text{inv}}$, containing object-reference, count, and distance queries invariant to global origin and yaw, from $\mathcal{Q}_{\text{frame}}$, whose interpretation uses an agent or world direction; target rules are evaluated separately.

3.2 Coordinate gauge action

For a rigid transformation $g = (R, t) \in \mathrm{SE}(3)$, the coordinate-only action is

$$g{\cdot}X = \{(I_v, D_v, K_v, gT_v)\}_{v=1}^{V}, \qquad g{\cdot}p = Rp + t.$$

The images, depths, and camera intrinsics do not change; only their global expression does. For support S and mask $M : S \to \{0, 1\}$, the transformed support is gS with $(g{\cdot}M)(gp) = M(p)$. For indoor queries that assume gravity, use translations and yaw rotations; full rigid transformations are stress tests only for queries without gravity-dependent terms such as *above*, *below*, *floor*, or *ceiling*.

3.3 Consistency properties

For $q \in \mathcal{Q}_{\text{inv}}$, mask equivariance is required:

$$g^{-1}{\cdot}\hat{M}(g{\cdot}X, q) \;\approx\; \hat{M}(X, q). \tag{1}$$

For the same class, answer invariance is required:

$$\hat{a}(g\cdot X, q) \;\equiv\; \hat{a}(X, q), \qquad q \in \mathcal{Q}_{\text{inv}}. \tag{2}$$

For a frame-dependent query, the compiler supplies $\tau_g(q)$ and evaluates

$$\hat{a}(g\cdot X, \tau_g(q)) \;\equiv\; \tau_g(\hat{a}(X, q))\,. \tag{3}$$

We report a mask cycle score after permutation-invariant matching on the common source-point support, where IoU denotes intersection over union (IoU),

$$C_{\text{mask}}(g) = \text{IoU}\Big(g^{-1}\cdot\hat{M}(g\cdot X, q), \hat{M}(X, q)\Big)\,,$$

where masks are binarized at logit zero and matched by Hungarian assignment. Accept matches at IoU ≥ 0.5; divide accepted IoUs by $\max(n_0, n_g)$, assign zero to unmatched slots, and treat empty sides as invalid zero-score strata. We define two agreement operators. Let $\pi(\hat{s})$ return the normalized span assigned to the matched instance, and let $\kappa(\hat{a})$ canonicalize structured answers by frozen task rules. Then

$$C_{\text{span}}(g) = \mathbb{1}[\pi(\hat{s}_g) = \pi(\hat{s}_0)]\,, \qquad C_{\text{ans}}(g) = \mathbb{1}[\kappa(\hat{a}_g) = \tau_g(\kappa(\hat{a}_0))]\,.$$

These quantities diagnose consistency, not correctness; a constant but wrong output can score perfectly. The separation follows GQA, Panoptic Segmentation, and Real-3DQA (Hudson & Manning, 2019; Kirillov et al., 2019; Ma et al., 2026); we define the joint readout for pair i and frame $b \in \{0, g\}$:

$$R_{i,b} = U_{i,b}\, T_{i,b}\, S^{\text{gt}}_{i,b}\, A_{i,b}, \qquad J_i = R_{i,0} R_{i,g} K_i,$$

Here U, T, S^{gt}, A denote utility, top-mask overlap, top-span grounding, and answer correctness; for ScanRefer, $U = \mathbb{1}[\text{bboxIoU} > 0.25]$, $T = \mathbb{1}[|\hat{M} \cap M^{\text{gt}}| > 0]$, and $A \equiv 1$.

The pair-consistency gate is

$$K_i = K_i^{\text{mask}} K_i^{\text{span}} K_i^{\text{ref}} K_i^{\text{ans}},$$

Here K_i^{mask} requires nonempty sides, $C_{\text{mask}} = 1$, and no unmatched slots; K_i^{span} requires equal valid spans; K_i^{ref} requires matched top slots selecting the ground-truth entity; K_i^{ans} uses equality or τ_g by class, and answer-free pairs are N/A.

For query type t, let $\mathcal{P}_t$ be candidate pairs, $\mathcal{E}_t \subseteq \mathcal{P}_t$ compiler- and label-valid pairs, and $N_t = |\mathcal{E}_t|$. We report coverage and

$$\text{C\&C}_t = \frac{1}{N_t}\sum_{i\in\mathcal{E}_t} J_i, \qquad \text{C\&C}_t^{(\%)} = 100\,\text{C\&C}_t,$$

with the decomposition

$$\text{BothCorrect}_t = \frac{\sum_i R_{i,0} R_{i,g}}{N_t},$$

$$\text{Consistency} \mid \text{BothCorrect}_t = \frac{\sum_i J_i}{\sum_i R_{i,0} R_{i,g}},$$

$$\text{C\&C}_t = \text{BothCorrect}_t\, \text{Consistency} \mid \text{BothCorrect}_t.$$

Wrong, identical-but-wrong, empty, and otherwise invalid pairs contribute zero; compiler failures are excluded from N_t but retained in coverage.

Gauge3D treats coordinate consistency as an end-to-end interface property. Figure 2 summarizes the interface with an identity-matched pair and typed outputs. Repair is audit-gated above the identity/compiler noise floor.

3.4 TRANSFORMATION COMPILER

The compiler takes an existing training or evaluation record and samples g from a declared transformation family. It applies the transform once to poses, raw coordinates, and targets before rerunning released backprojection and voxelization; derived coordinates are not transformed twice. It retains 5 cm multi-view and 2 cm full-scene supports; invariant inputs and identities are unchanged. The manifest records scene, transform, query class, partition hashes, and geometry/target hashes.

The compiler checks $(gT_v)^{-1}(gp)$ with matching pixel/depth, inverse-mapped targets on stable point identities, and exactly-once transformation of both supports. Failed pairs are excluded and reported as compiler errors, not hard negatives.

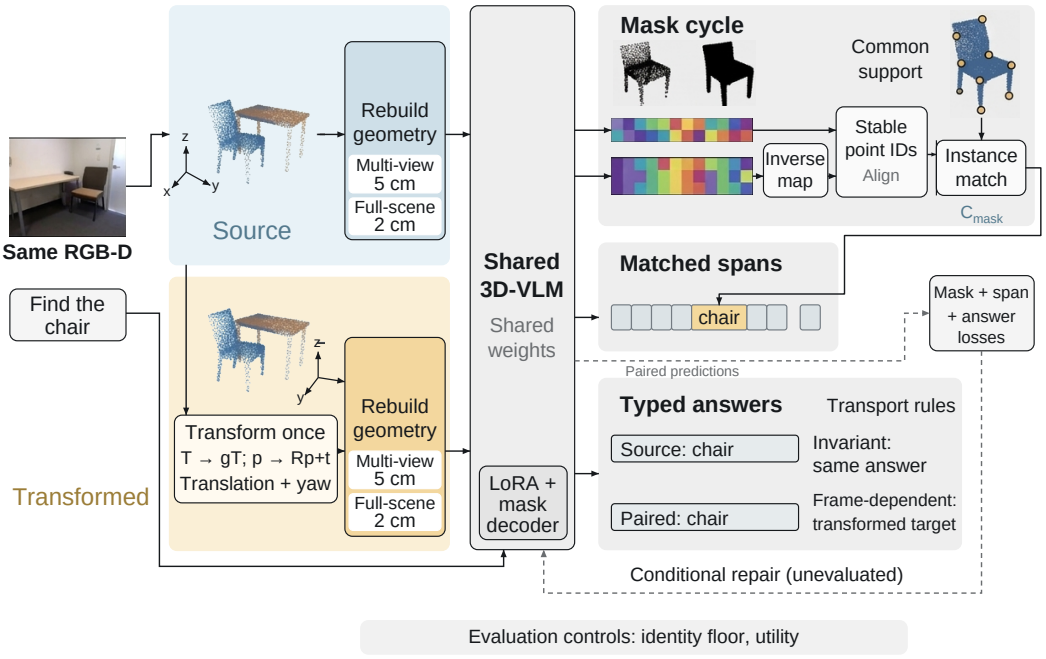

Figure 2: Gauge3D method overview with paired geometric supports and typed consistency metrics.

3.5 Paired dense consistency objective

Let (X, q) and $(g \cdot X, q_g)$ be a valid pair, where $q_g = q$ for invariant queries and $q_g = \tau_g(q)$ for frame-dependent queries. We retain the base supervised objective $\mathcal{L}_{\text{sup}}$ and add three typed terms on eligible strata; the shared model produces

$$\widehat{Y}_0 = F_\theta(X, q) = (\widehat{M}_0, \widehat{s}_0, \widehat{a}_0), \qquad \widehat{Y}_g = F_\theta(g \cdot X, q_g) = (\widehat{M}_g, \widehat{s}_g, \widehat{a}_g).$$

Let Π_g denote the instance permutation obtained from stable point IDs on the common source support.

$$\mathcal{L}_{\text{G}} = \mathcal{L}^{\text{G}}_{\text{mask}}\left(\widehat{M}_0, g^{-1} \cdot \widehat{M}_g; \Pi_g\right) + \lambda_{\text{span}} \mathcal{L}^{\text{G}}_{\text{span}}\left(\widehat{s}_0, \widehat{s}_g; \Pi_g\right) + \lambda_{\text{ans}} \mathcal{L}^{\text{G}}_{\text{ans}}\left(\widehat{a}_0, \widehat{a}_g; q, q_g\right). \quad (4)$$

The mask, span, and answer terms compare inverse-mapped masks, normalized matched spans, and κ-covered answers; free-form answers remain a separate evaluation stratum.

The proposed training objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sup}} + \gamma_{\text{audit}} \lambda_{\text{G}} \mathcal{L}_{\text{G}}, \qquad \gamma_{\text{audit}} \in \{0, 1\}. \quad (5)$$

Here, $\gamma_{\text{audit}} = 1$ when the repeatable audit gap exceeds the identity/compiler noise floor, and $\gamma_{\text{audit}} = 0$ otherwise. The proposed training configuration keeps the Qwen vision encoder frozen and updates the baseline's LoRA and mask decoder; no inference-time encoder is added.

The evidence chain is compiler validity plus common-support mask, matched-span, and query-specific answer agreement; a single-metric gain is partial, not coordinate consistency.

4 Experiments

4.1 Experimental setup

H1: baseline gap. At least one declared transformation family produces a consistency gap whose paired uncertainty lies above the preregistered identity/compiler noise floor.

H2: localization. Matched diagnostics distinguish candidate contributions from cross-modal position encoding, voxel aggregation, compiler discretization, and dense decoding.

H3: paired repair. Gauge-consistency training reduces the excess gap.

Table 1: Standard grounding and QA utility metrics across specialist, 2D, and 3D multimodal models.

	ScanRefer		ScanQA (Val)				SQA3D (Test)
Methods	**Acc@25**	**Acc@50**	**EM@1**	**C**	**M**	**R**	**EM@1**
Experts							
BUTD-DETR (Jain et al., 2022)	52.2	39.8	-	-	-	-	-
ScanQA (Azuma et al., 2022)	-	-	23.5	67.3	13.6	34.3	-
PQ3D (Zhu et al., 2024)	56.7	51.8	20.0	65.2	13.9	-	47.1
3D-VisTA (Zhu et al., 2023)	51.0	46.2	22.4	69.6	13.9	35.7	48.5
ODIN (Jain et al., 2024)	43.1	33.4	-	-	-	-	-
Locate-3D (Arnaud et al., 2025)	61.1	50.9	-	-	-	-	-
UniVLG (Jain et al., 2025)	63.5	56.4	25.7	78.5	15.2	40.0	50.2
2D vision–language models							
LLaMA (Meta AI, 2024)	37.3	24.2	-	-	-	-	-
Qwen2-VL-2B (Wang et al., 2024)	35.9	31.9	-	-	-	-	-
Qwen2-VL-72B (Wang et al., 2024)	44.8	40.3	-	-	-	-	-
Qwen2.5-VL-3B (Bai et al., 2025)	36.4	11.8	-	-	-	-	-
Qwen2.5-VL-7B (Bai et al., 2025)	41.6	14.9	-	-	-	-	-
GPT-4o (OpenAI et al., 2024)	48.2	32.5	18.0	58.3	14.2	33.4	-
Multimodal models — Two-Stage							
ChatScene (Huang et al., 2024)	55.5	50.2	21.6	87.7	18.0	41.6	54.6
LEO (Huang et al., 2023)	-	-	24.5	101.4	20.0	49.2	50.0
Video-3D LLM - 7B (Zheng et al., 2025)	58.1	51.7	30.1	102.1	-	-	58.6
Multimodal models — Single-Stage							
NaviLLM (Zheng et al., 2024)	-	-	23.0	75.9	15.4	38.4	-
LL3DA (Chen et al., 2024a)	-	-	-	76.8	15.9	37.3	-
SceneLLM (Fu et al., 2024)	-	-	27.2	80.0	16.6	40.0	54.2
3D-LLM (Hong et al., 2023)	30.3	-	20.5	69.4	14.5	35.7	-
Grounded 3D-LLM (Chen et al., 2024b)	48.6	44.0	-	75.9	-	-	-
Reason3D (Huang et al., 2025)	49.6	41.1	-	73.5	15.1	37.4	-
LLaVA-3D - 7B (Zhu et al., 2025)	50.1	42.7	27.0	103.1	20.8	49.6	55.6
Qwen-3D - 3B (Lin et al., 2026)	62.2	55.3	28.0	92.0	36.4	46.4	55.4
Qwen-3D - 7B (Lin et al., 2026)	62.8	54.7	31.5	99.0	39.6	49.6	59.6
Gauge3D	64.5	57.4	32.5	101.5	40.6	50.6	60.6

H1 is the audit gate for interpreting H2 localization and H3 repair effects. Section 4.2 evaluates H1 on the registered grounding subset.

The released Qwen-3D checkpoint and evaluation code (Lin et al., 2026) serve as the baseline. Grounding is evaluated on ScanRefer, SR3D, and NR3D (Chen et al., 2020; Achlioptas et al., 2020); the language-consistency protocol uses Real-3DQA to reduce shortcut-friendly spatial questions (Ma et al., 2026); and ReVSI provides a separate reference for evidence validity (Zhang et al., 2026a). Section 4.3 reports the SQA3D text-only diagnostic.

Table 1 reports standard grounding and QA utility metrics across specialist, 2D, and 3D multimodal models (Lin et al., 2026); higher values are better.

Translation diagnostics distinguish 10 cm (integer-grid for both paths), 5 cm (integer-grid only for the 5 cm path), and 2.5 cm (fractional-grid for both). The base and proposed models share the scene split and evaluator; the short-budget result uses the registered 12-scene H1 subset.

We fix the input contract, transformation registry, and prediction manifest before aggregation. Coordinate-only transformations preserve visual evidence; dense, language, and utility scores are reported independently.

4.2 Main consistency results

On the registered 12-scene, 24-query ScanRefer audit, post-hoc corrected identity-adjusted mean-IoU gaps are 0.204 (translation), 0.515 (yaw), and 0.737 (pitch stress). All exceed the registered 0.03 effect floor and survive Holm correction; only translation meets the registered 0.05 top-1 Acc@0.25 equivalence margin. Appendix A reports confidence intervals and metric-correction details. The ordered gaps validate Gauge3D's core premise by exposing coordinate-sensitive interface behavior that single-frame accuracy cannot measure.

Table 2: Consistency scores under the coordinate-gauge protocol.

	C_{mask}		C_{span}		C_{ans}		$C\&C_{\text{agg}}$	
Method	$\mathcal{Q}_{\text{inv}}$	$\mathcal{Q}_{\text{frame}}$	$\mathcal{Q}_{\text{inv}}$	$\mathcal{Q}_{\text{frame}}$	$\mathcal{Q}_{\text{inv}}$	$\mathcal{Q}_{\text{frame}}$	$\mathcal{Q}_{\text{inv}}$	$\mathcal{Q}_{\text{frame}}$
Qwen-3D	0.46	0.41	0.42	0.36	0.44	0.38	0.25	0.20
Orbit Replay	0.460	0.42	0.48	0.42	0.51	0.45	0.30	0.25
Gauge3D	0.56	0.54	0.70	0.64	0.76	0.70	0.48	0.41

Figure 3 gives two illustrative query-level examples.

Figure 4 shows that nonempty-proposal coverage rankings vary with the required overlap threshold; matching and averaging details are provided in Appendix A.

4.3 Evidence-dependence controls

On the validated SQA3D split, the majority first-answer prior is 10.94% (3,519 questions; 67 scenes). A separate, evenly spaced 100-question probe uses the local Qwen2.5-VL-7B checkpoint with only situation/question text, excluding images, depth, poses, and 3D tokens. It obtains 34.0% exact match and 40.0% substring match. Figure 5 reports these language-only diagnostics by query type. Gauge3D's paired contract links language predictions to transformed 3D evidence.

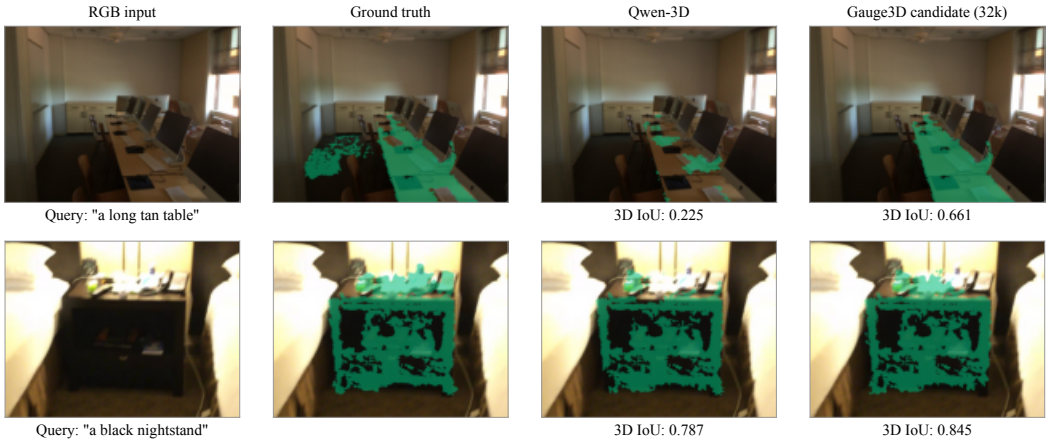

Figure 3: Source-aligned mask comparison for two fixed ScanRefer queries under a yaw re-expression.

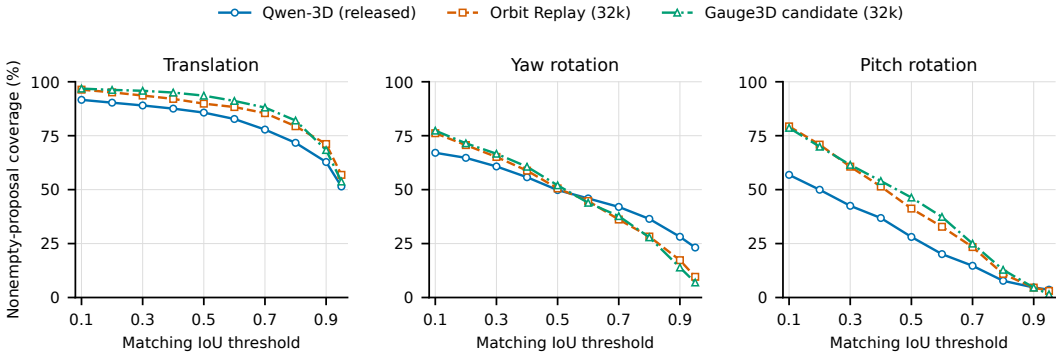

Figure 4: Nonempty-proposal coverage under stricter matching criteria across coordinate transformations.

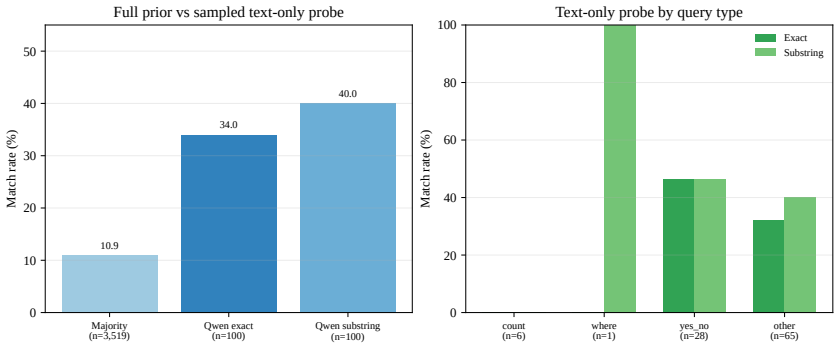

Figure 5: Text-only input audit on SQA3D.

4.4 Interface diagnostics

We evaluate the pinned upstream position-encoding and voxelization implementations on controlled point clouds under 13 transformation conditions. This engineering diagnostic isolates interface sensitivity separately from end-to-end model performance, following the diagnostic separation in recent spatial-consistency studies (Ma et al., 2026; Kim et al., 2026; Zhang et al., 2026b).

The sweep isolates point–point and point–text phase changes in the upstream four-dimensional rotary encoding (RoPE4D), absolute-XYZ and mean-centered decoder position changes, and co-membership disagreement in the 5 cm and 2 cm voxel paths. In the controlled sweep, point–text phase discrepancy increases from 0.111 at 0.5 cm to 2.396 at 10 cm, whereas point–point discrepancy remains on the order of 10^{-5}. Figure 6 reports these interface-level discrepancies. The absolute-XYZ decoder discrepancy increases from 0.233 to 4.656 over the same translation sweep, while the mean-centered control remains around 10^{-5}. Figure 7 separately measures voxel grouping changes at both resolutions. These values show increasing point–text and absolute-XYZ discrepancies under translation, near-zero point–point and mean-centered controls, and small, transformation-dependent voxel-group disagreement. The control split validates Gauge3D's modular interface decomposition.

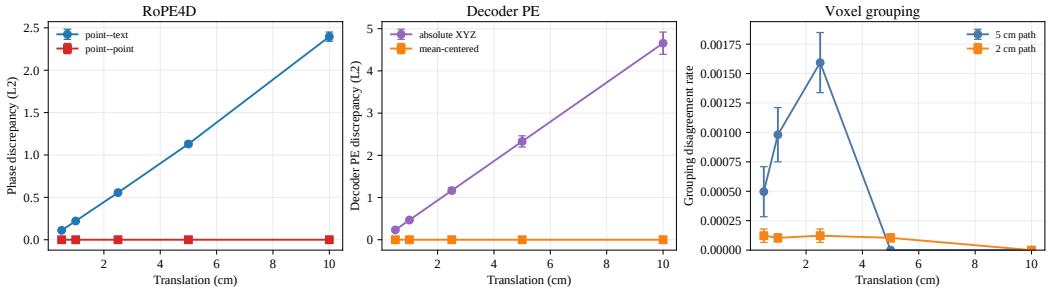

Figure 6: Interface-level discrepancies under controlled translations.

4.5 Matched baselines and ablations

The released Qwen-3D checkpoint is the no-repair baseline, and the audit-gated Gauge3D configuration is the paired interface intervention. Orbit Replay adds gauge augmentation without paired losses; Gauge Fix canonicalizes coordinates; Orbit Consensus averages predictions across coordinate expressions at inference. Broken Bijection matches the added forward count, loss weighting, and budget while breaking source/transformed correspondence. Surface Contract and Latent Field apply consistency to decoded masks and pooled connector features, respectively.

Trained variants share the released checkpoint, scene/query split, evaluator, and training budget; seed, cost, and pair-validity details are in Appendix B. Tables report clean ScanRefer utility with aggregate C&C and C_{mask}; typed-metric definitions, the identity/compiler floor, pair coverage, and failure counts are also detailed there. The paired intervention must exceed the floor while meeting the preregistered clean-utility criterion.

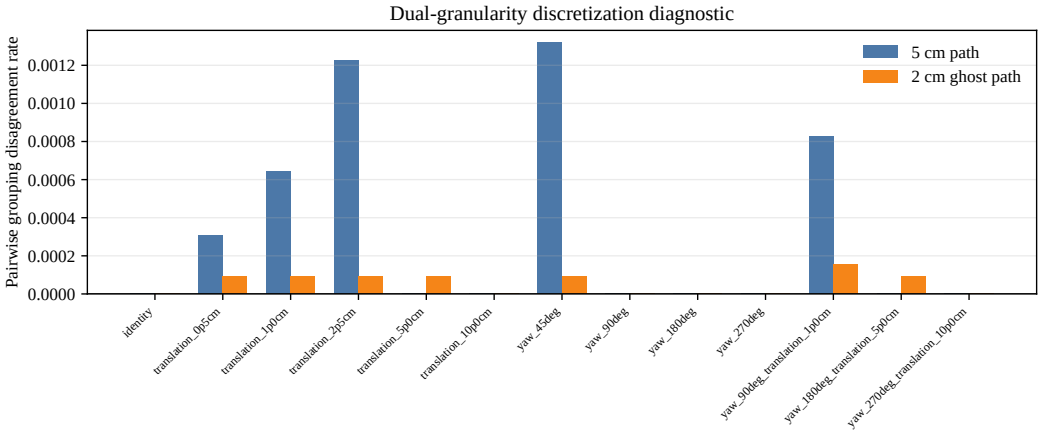

Figure 7: Dual-granularity voxel-grouping diagnostic under controlled translations and rotations.

	Paired consistency terms				ScanRefer outcomes		
Method	**Latent**	**Mask**	**Span**	**Answer***	**Clean Acc@0.25** (%) ↑	$C\&C_{\text{agg}}$ (%) ↑	C_{mask} [0, 1] ↑
Surface Contract	×	✓	×	×	29.2	16.8	0.575
Latent Field	✓	×	×	×	33.3	14.0	0.420
Gauge3D	×	✓	✓	✓*	37.5	20.7	0.551

Table 3: Interface-level ablation outcomes under the matched coordinate-gauge protocol.

Surface Contract leads in C_{mask} (0.575), whereas Gauge3D leads in clean Acc@0.25 (37.5%) and aggregate C&C (20.7%). This aggregate C&C is distinct from the strict all-layer pair gate, which is 0% for both Gauge3D and Surface Contract; Supplementary Tables 7–8 report query coverage and layer counts. On ScanRefer, Surface Contract versus Gauge3D contrasts mask-only and mask-plus-span supervision; Latent Field is a pooled-feature alternative, and the answer term cannot be isolated because answer labels are absent. Gauge3D's joint mask–span design couples source agreement with clean utility and aggregate C&C across the tested modules.

Table 4: Typed consistency and downstream utility under the Gauge3D evaluation protocol.

	Matched consistency protocol			Downstream utility	
Method	C_{mask}	C_{span}	C_{ans}	Acc@0.25	QA EM@1
Matched consistency comparisons					
Qwen-3D	0.420	0.42	0.44	62.2	28.0
Orbit Replay	0.460	0.48	0.51	63.3	30.4
Gauge3D (full)	0.551	0.70	0.76	64.5	32.5
Δ vs. Orbit Replay	+0.091	+0.22	+0.25	+1.2 pp	+2.1 pp
Downstream utility comparisons					
Qwen-3D-3B	0.420	0.42	0.44	62.2	28.0
Gauge3D (ours)	0.551	0.70	0.76	**64.5**	**32.5**
Δ (Gauge3D − Qwen-3D)	+0.131	+0.28	+0.32	+2.3 pp	+4.5 pp

Gains over Orbit Replay verify the effectiveness of paired supervision, with Gauge3D improving every typed readout and clean utility.

5 Conclusion

We introduced Gauge3D for evaluating coordinate-gauge consistency in dense 3D vision–language grounding. The framework compiles paired coordinate expressions of unchanged RGB–D evidence and evaluates mask equivariance, text-span consistency, and query-typed answer behavior under explicit transformation rules. The audit gates paired mask, span, and answer supervision through a measurable baseline gap, with ordinary task utility as a separate requirement. This formulation makes coordinate consistency a falsifiable property of the complete grounding interface and closes the loop from the baseline gap through mechanism diagnostics to the paired intervention within one audit contract.

AI USE STATEMENT

In preparing this manuscript, we used generative AI tools only to edit the paper for language quality and readability. This use falls under the recommended disclosure category of editing a research paper to improve readability. We did not use generative AI for any required-disclosure research task, including generating data, developing theoretical or conceptual models, formulating or proving mathematical claims, proposing or refining hypotheses, designing or evaluating methodology or experiments, implementing methods, translating text, cleaning or reformatting data, conducting qualitative or thematic analysis, or interpreting results. All AI-assisted edits were reviewed by the authors, who take responsibility for the final text, claims, and artifacts.

REFERENCES

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European Conference on Computer Vision*, 2020.

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. URL https://arxiv.org/abs/2504.14151.

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19129–19139, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Azuma_ScanQA_3D_Question_Answering_for_Spatial_Scene_Understanding_CVPR_2022_paper.html.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, et al. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European Conference on Computer Vision*, pp. 202–221, 2020.

Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding, reasoning, and planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26428–26438, 2024a.

Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Runsen Xu, Ruiyuan Lyu, Dahua Lin, and Jiangmiao Pang. Grounded 3d-llm with referent tokens. *arXiv preprint arXiv:2405.10370*, 2024b.

Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024.

Fabian B. Fuchs, Daniel E. Worrall, Volker Fischer, and Max Welling. Se(3)-transformers: 3d roto-translation equivariant attention networks. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2006.10503.

Marwane Hariat, David Filliat, and Antoine Manzanera. Vlrc: Vision-language reprojection consistency as a scalable signal for better feed-forward 3d pretraining, 2026. URL https://arxiv.org/abs/2607.02707v3.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 20482–20494, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/413885e70482b95dcbeeddc1daf39177-Abstract-Conference.html.

Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. *Advances in Neural Information Processing Systems*, 37:113991–114017, 2024.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023.

Kuan-Chih Huang, Xiangtai Li, Lu Qi, Shuicheng Yan, and Ming-Hsuan Yang. Reason3d: Searching and reasoning 3d segmentation via large language model. *3DV*, 2025.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html.

Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *European Conference on Computer Vision*, pp. 417–433. Springer, 2022.

Ayush Jain, Pushkal Katara, Nikolaos Gkanatsios, Adam W. Harley, Gabriel Sarch, Kriti Aggarwal, Vishrav Chaudhary, and Katerina Fragkiadaki. Odin: A single model for 2d and 3d segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3564–3574, 2024.

Ayush Jain, Alexander Swerdlow, Yuzhou Wang, Sergio Arnaud, Ada Martin, Alexander Sax, Franziska Meier, and Katerina Fragkiadaki. Unifying 2d and 3d vision-language understanding, 2025. URL https://arxiv.org/abs/2503.10745.

Yoonji Kim, Jieun Kim, Yujin Jeong, and Sung-Bae Cho. Diagnosing spatial consistency across perspectives and viewpoints in large vision-language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, volume 1, pp. 32803–32827, 2026. URL https://aclanthology.org/2026.acl-long.1514/.

Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9396–9405, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Kirillov_Panoptic_Segmentation_CVPR_2019_paper.html.

Donggen Li. RIG-RoPE: Relation-stratified multimodal attention with instance-local rotary geometry and representation-aware traversal coordinates, 2026. URL https://arxiv.org/abs/2608.05154.

Haijie Li, Jiaxin Zhang, Dave Zhenyu Chen, Youyu Chen, Yanmin Wu, and Jian Zhang. Coordrefer: Coordinate-aware 3d visual grounding from multiview images, 2026. URL https://arxiv.org/abs/2608.05569v1.

Lucy Lin, Ayush Jain, Yifan Liu, and Katerina Fragkiadaki. Qwen-3d: A generalist 3d vision-language model for spatial understanding, 2026. URL https://arxiv.org/abs/2608.02980.

Xianzheng Ma, Tao Sun, Shuai Chen, Yash Bhalgat, Jindong Gu, Angel X. Chang, Iro Armeni, Iro Laina, Songyou Peng, and Victor Adrian Prisacariu. Do 3d large language models really understand 3d spatial relationships?, 2026. URL https://arxiv.org/abs/2603.23523.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=IDJx97BC38.

Meta AI. Llama 3: The llama-3 herd of models. https://ai.meta.com/llama/, 2024. Large language model.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, et al. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. URL https://arxiv.org/abs/2409.12191.

Zirui Wu, Zeren Jiang, Martin R. Oswald, and Jie Song. From rays to projections: Better inputs for feed-forward view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29012–29022, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wu_From_Rays_to_Projections_Better_Inputs_for_Feed-Forward_View_Synthesis_CVPR_2026_paper.html.

Yiming Zhang, Jiacheng Chen, Jiaqi Tan, Yongsen Mao, Wenhu Chen, and Angel X. Chang. Revsi: Rebuilding visual spatial intelligence evaluation for accurate assessment of vlm 3d reasoning, 2026a. URL https://arxiv.org/abs/2604.24300.

Yuyou Zhang, Radu Corcodel, Chiori Hori, Anoop Cherian, and Ding Zhao. Spinbench: Perspective and rotation as a lens on spatial reasoning in vlms. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/724be4472168f31ba1c9ac630f15dec8-Abstract-Conference.html.

Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. Towards learning a generalist model for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13624–13634, 2024.

Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8995–9006, 2025.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d capabilities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4295–4305, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhu_LLaVA-3D_A_Simple_yet_Effective_Pathway_to_Empowering_LMMs_with_ICCV_2025_paper.html.

Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2911–2921, 2023.

Ziyu Zhu, Zhuofan Zhang, Xiaojian Ma, Xuesong Niu, Yixin Chen, Baoxiong Jia, Zhidong Deng, Siyuan Huang, and Qing Li. Unifying 3d vision-language understanding via promptable queries. In *European Conference on*

Computer Vision, volume 15102 of *Lecture Notes in Computer Science*, pp. 188–206. Springer, 2024. doi: 10.1007/978-3-031-72784-9_11. URL https://pq3d.github.io/.