PoseRefer: Pathway-Local Parameters for Semantically Grounded Reference Resolution

Anna Deichler
KTH Royal Institute of Technology, Stockholm, Sweden
deichler@kth.se

***Abstract*—A robot resolving "put the cup on that one" must fuse gesture, language, and scene geometry, yet 3D grounding benchmarks only partially capture this regime: descriptions are written post-hoc, gestures are templated, or pointing is staged for the camera. MM-Conv captures natural co-speech gesture from dyadic VR interaction alongside full-body motion capture and 3D scene graphs. We use it to evaluate pose–language fusion with a *decoupled late-fusion* architecture in which pose and text pathways share no learned parameters. The two choices together make category, pose, and text contributions easier to isolate through controlled ablations. Fusion with frozen MiniLM category embeddings exceeds pose alone and the best text-only pathway on every reference type, reaching 31.9% top-1. The learned scalar gate flips between opposing policies depending on whether the text pathway has category access. This is a reliability diagnostic: fusion-accuracy claims for semantic grounding systems are indistinguishable from category-representation artifacts unless pathways are architecturally decoupled.**

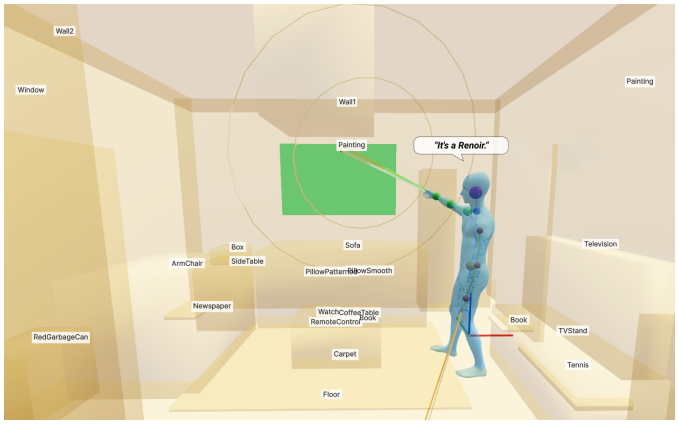

Fig. 1. PoseRefer in a 3D scene. The SMPL-X mesh shows the speaker's pose; the pointing ray (green) and head direction (magenta) define angular proximity to candidate objects (bounding boxes).

I. Introduction

When a person points at a chair and says "put the cup on that one," a service robot observing from a third-person view must integrate gesture, language, and scene geometry to identify the referent. This problem, exocentric reference resolution, has received less attention than its 2D image-grounded cousin, despite being a regime central to many household and assistive robotics scenarios.

Existing 3D grounding benchmarks substitute one or more of the naturalistic ingredients: post-hoc descriptions on static scans (ScanRefer [1], Sr3D/Nr3D [2]), template-generated language with synthetic 2D pose (Ges3ViG [3]), or single-user pointing performed for the camera (YouRefIt [4]). MM-Conv [5] differs in kind: references arose from dyadic VR interaction where two participants worked through a referential task in a shared 3D scene, with full-body motion capture, word-level speech alignment, and ground-truth scene graphs. It contains $\sim$4,000 references across 5 rooms, with $\sim 55\%$ accompanied by a detectable pointing or pointing-variant gesture and $\sim 45\%$ without. Reliable grounding of such references is a prerequisite for any downstream action selection that depends on semantic scene understanding.

On naturalistic gesture data, two confounds become visible that synthetic data hides. *First*, per-object category embeddings entangle with all three pathways, pose, text, and the fusion gate, through shared encoders and shared parameter tables. Ablating category in such architectures simultaneously alters every signal and every learned representation, making fusion comparisons hard to interpret. *Second*, the representation chosen for those per-object categories matters substantially: in our experiments, a learned 16-dimensional category embedding is insufficient at MM-Conv's per-category training counts, and the fusion gate inherits this noise. Both problems are invisible unless the architecture allows direct ablation of individual signals *and* the category representation itself is varied.

Contributions.

1) *Position:* naturalistic referential gesture data is an important testbed for claims about semantic grounding in multimodal perception. Nearly half of such references occur without clear pointing, a regime synthetic benchmarks largely exclude.
2) *Method:* a decoupled late-fusion architecture with *pathway-local* category embedding tables (no shared parameters between pose and text), and frozen pretrained semantic embeddings for per-object category on the text pathway. The two choices together reduce category, pose, and text to controllable ablations.
3) *Empirical findings:* on MM-Conv, (a) the fusion gate commits to opposing policies depending on whether category is available on the text pathway ($\alpha = 0.84 \rightarrow 0.22$), a diagnostic reproducible across cross-validation folds; (b) pose and text are asymmetrically complementary: pose dominates pointing tiers, text dominates non-pointing tiers; (c) with pathway-local parameters

and frozen MiniLM category embeddings, scalar fusion exceeds both singletons across every reference type, reaching 31.9% aggregate top-1, a gain of 13.1 points over pose alone and 6.9 over the strongest text-only configuration.

4) *Reliability implication:* fusion gains in architectures with shared encoders or shared category embeddings may reflect the category-representation choice rather than properties of fusion itself. Our decoupled setup makes this directly testable and motivates reporting both learned and frozen category representations when evaluating multimodal fusion for semantic grounding.

II. Related Work

3D visual grounding operates on post-hoc scene descriptions without the speaker's body (ScanRefer [1], Sr3D/Nr3D [2], BUTD-DETR [6], EDA [7]); Ges3ViG [3] adds template language with synthetic 2D pose and YouRefIt [4] captures real but single-user 2D reference. 3D scene graphs have emerged as a compact semantic environment representation for embodied tasks [8], [9]; MM-Conv's ground-truth scene graphs let us isolate pose–language fusion from scene-perception error. Wang et al. [10] use a distance field between skeleton joints and scene points for language-guided motion generation; we adapt the idea to angular distance over discrete candidates. Deichler et al. [11] generate context-aware pointing gestures for embodied agents; we address the inverse problem of interpreting observed gestures for reference resolution. Related embodied fusion work explores task-adaptive gated routing [12] and exocentric reference benchmarks [13]. Frozen pretrained language spaces appear for semantic matching in open-vocabulary pipelines [14], but typically as part of a broader vision–language stack rather than an isolated per-object handle in a dedicated fusion pathway.

III. Dataset and Setup

MM-Conv contains approximately 4,200 referring expressions from 6.7 h of dyadic VR interaction across 5 AI2-THOR rooms with 42–61 candidate objects per room, synchronized speech (word-level timestamps), full-body SMPL-X parameters at 30 fps, and 3D scene graphs. Expressions are 38% exact noun phrases, 14% partitives, 48% pronominals (e.g.,"that one", "this one"). After filtering for valid targets, text features, and valid temporal windows we evaluate on 3,764 references. We adopt a third-person robot camera view; scene-graph ground truth is unchanged.

Category vocabulary. Scene-graph labels are lowercased, compound labels split, and color/material/size modifiers stripped ("RedVase" → "vase"), addressing training-signal sparsity in the raw vocabulary. The normalized vocabulary contains 50 categories. Stripping discards potentially grounding-relevant attributes (e.g., color), a systematic ablation of modifier retention is deferred to future work.

Gesture classification. A random-forest classifier on 16 SMPL-X kinematic features separates clear pointing from no-gesture (AUC 0.97). Classifier scores stratify references into four confidence tiers, T1–T4, plus a post-hoc T5 for two-arm pointing; we use the binary partition *pointing* (T1∪T2∪T5, $n = 2{,}068$, 55%) vs. *non-pointing* (T3∪T4, $n = 1{,}696$) throughout. Tiers are descriptive only, never used as model input. Per-tier counts are reported in Table II.

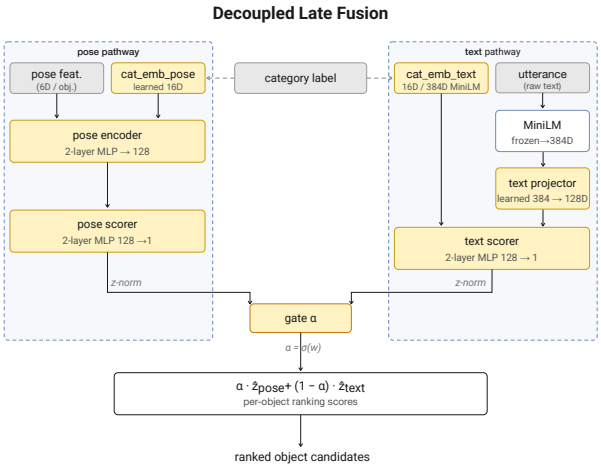

Fig. 2. PoseRefer architecture. Pose and text pathways operate independently with no shared learned parameters. The text pathway encodes the utterance via a frozen MiniLM-L6 model (→384D) before a learned linear projection to 128D. Each pathway has its own per-object category embedding table: learned 16D on the pose side; on the text side, either learned 16D or (default in $\text{PT}_{\text{minilm}}$) a frozen MiniLM encoding of the canonical category name. Per-pathway scores are z-normalized and combined by a learned scalar gate α.

IV. Method: Decoupled Late Fusion

We score each candidate object by its angular proximity to the speaker's body channels over a temporal window aligned to the utterance, and, the key architectural choice of this paper, route that score through a pathway that is *architecturally disjoint* from the text pathway and shares no learned parameters with it.

Temporal angular affordance. For object n and channel $j \in \{\text{R arm}, \text{L arm}, \text{head}, \text{body}\}$, the per-frame angle is $\theta_{n,j}^{(t)} = \angle(\mathbf{d}_j^{(t)}, \mathbf{p}_n - \mathbf{q}_j^{(t)})$, where $\mathbf{d}_j^{(t)}$ is the channel direction (pointing ray for arms, facing for head/body), $\mathbf{q}_j^{(t)}$ its origin (wrist, eye midpoint, pelvis), and $\mathbf{p}_n$ the object centroid. Angles map to scores via a Gaussian kernel $s_{n,j}^{(t)} = \exp(-\theta^2/2\sigma_j^2)$ with $\sigma_{\text{arm}} = 15°$, $\sigma_{\text{head}} = 30°$, $\sigma_{\text{body}} = 45°$.

Temporal pooling. Arm scores are computed on a narrow ±10-frame window centered on the utterance hold frame (pointing is near-instantaneous and the hold annotation is accurate); head and body scores use a wider window spanning phrase-start-0.5 s to phrase-end$+0.5$ s (gaze and torso orientation persist through the utterance). We pool with both max and mean for arms (peak vs. sustained), max only for head (gaze is spiky), mean only for body (torso is stable), giving a 6-dimensional pose feature vector per object. Arm-extension gating was explored but did not improve pose-pathway accuracy and is not used in the reported results.

Decoupled pathways. Each candidate object receives scores from two architecturally separate pathways:

- *Pose pathway*: the 6-D angular affordance features, concatenated with a pose-local 16-D learned category embedding E_{pose} (when enabled), are fed to a 2-layer encoder $(6{+}|E_{\text{pose}}| \to 128 \to 128)$ followed by a 2-layer scorer $(128 \to 128 \to 1)$ producing a pose score s_n^{pose}.
- *Text pathway*: receives no world-frame geometry, isolating utterance semantics from spatial cues. A frozen MiniLM-L6 [15] utterance embedding (384-D) is projected to 128-D, broadcast across objects, concatenated per-object with a text-local category embedding E_{text} (when enabled), and passed through a 2-layer scorer $(128{+}|E_{\text{text}}| \to 128 \to 1)$ producing a text score s_n^{text}. Per-object, E_{text} is either a learned 16-D vector or a frozen 384-D MiniLM encoding of the category's canonical name; we ablate both choices.

The pathways share no learned parameters: neither MLP weights nor (when learned) category embedding tables.[1] Under a shared encoder, or a decoupled encoder with *shared* category embedding, ablating category alters every downstream signal at once and training-time gradients from one pathway shape the embedding seen by the other. Pathway-local embeddings turn "what does category contribute in each pathway?" into a controlled experiment.

Category representation on the text pathway. Under a learned 16-D embedding, each category must learn its representation from LORO-fold (Leave-one-room-out) training examples alone: with 50 categories and fold training sizes around 3,000, many categories have only a handful of training samples, producing unreliable vectors for the less-frequent labels. A frozen MiniLM encoding of the category name sidesteps this: categories are represented in the same 384-D semantic space as the utterance embedding, inheriting "vase"–"chair"–"table" similarity structure from the pretraining distribution. We treat the choice between learned and frozen MiniLM category embeddings as a methodological ablation; §V-C shows it is also the difference between 28.2% and 31.9% aggregate top-1 under fusion.

Fusion. Per-pathway scores are z-normalized per sample and combined with a learned global gate

$$s_n = \alpha\, \hat{s}_n^{\text{pose}} + (1-\alpha)\, \hat{s}_n^{\text{text}}, \qquad \alpha = \sigma(w),$$

with w initialized to 0 so that $\alpha = 0.5$ at the start of training. α is a single learned scalar per training run, shared across all samples.

Training. LORO cross-validation across the 5 rooms; 50 epochs, AdamW, lr 10^{-3}, batch 32, cosine schedule; cross-entropy over candidates within each sample. All numbers are reported as mean±std across three random seeds under LORO cross-validation; per-fold α values are listed where relevant. Dropout 0.3 in all MLPs; weight decay 10^{-4}. Paired t-tests ($n = 3$) compare seed-averaged LORO fold means.

[1]Under frozen-MiniLM-on-both, the two pathways read from the same constant buffer, a shared input, not a shared parameter.

TABLE I
REFERENCE RESOLUTION ACCURACY UNDER LEAVE-ONE-ROOM-OUT CROSS-VALIDATION. MEAN±STD OVER THREE RANDOM SEEDS.

Config	Pose	Text	Cat	Top-1	Top-5	α
Random	,	,	,	∼2.0	∼10.0	,
P	✓	,	,	18.8±0.3	51.7±0.3	,
P_{cat}	✓	,	L pose	18.9±0.6	49.3±0.9	,
T	,	✓	L text	22.5±2.4	49.2±0.4	,
T_{minilm}	,	✓	M text	25.0±0.4	50.5±0.4	,
PT_{nocat}	✓	✓	,	18.9±0.1	51.7±0.2	0.837±0.000
PT	✓	✓	L both	28.2±0.2	57.8±0.9	0.222±0.001
PT_{minilm}	✓	✓	M text	**31.9**±0.1	**60.2**±0.5	0.248±0.007
$PT_{minilm,both}$	✓	✓	M both	31.6±0.1	59.7±0.4	0.267±0.002

V. EMPIRICAL FINDINGS

Table I reports aggregate accuracy for eight configurations spanning combinations of pathway inclusion and category representation; Tables II and III stratify the headline configurations by gesture tier and reference type.

Three findings structure the analysis. First (§V-A), the learned fusion gate flips between opposing policies depending solely on whether the text pathway has a per-object category handle, a diagnostic that the pathway-local decoupling makes directly observable. Second (§V-B), pose and text are strongly asymmetrically complementary across gesture tiers, establishing the disagreement pattern fusion must exploit. Third (§V-C), with frozen MiniLM category embeddings on the text pathway, scalar fusion exceeds both singletons across every reference type, reaching 31.9% top-1.

A. The Fusion Gate Flips with Category Presence

The learned global gate commits to opposing policies depending on whether the text pathway has access to a per-object category handle. Under PT_{nocat} (no category on either pathway), the gate settles at $\alpha = 0.837$ across folds, trusting pose heavily. Under PT (learned category embeddings on both pathways), the same architecture on the same 3,764 references under the same LORO folds settles at $\alpha = 0.222$, nearly the opposite policy. Only the text-pathway category handle differs between the two runs; every other hyperparameter, random seed, and fold assignment is held fixed.

The gate does not "drift" to its final α; its final value is highly reproducible across the five LORO folds (per-fold α: PT_{nocat} in $[0.831, 0.842]$; PT in $[0.194, 0.237]$). This is a deterministic architectural response, not a training-noise effect. The interpretation is mechanical: when the text pathway has no per-object distinguishing signal beyond a broadcast utterance embedding, architecturally the case here, since the text scorer's input text_projector(utterance) is identical across all candidate objects within a sample, the gate correctly discounts it and trusts pose. When category makes the text pathway per-object-informative, the gate correctly shifts its weight to text.

This has two consequences. Scalar fusion gates are sensitive to architectural choices that determine whether each pathway carries per-object signal: small representation changes produce

TABLE II
TOP-1 ACCURACY BY GESTURE TIER. MEAN±STD OVER THREE RANDOM SEEDS UNDER LORO CROSS-VALIDATION.

Config	T1	T2	T3	T4	T5
n	1,344	558	153	1,543	166
P	31.5±0.6	20.4±0.7	9.6±0.8	7.5±0.3	24.3±0.3
T_{minilm}	22.8±0.5	25.0±0.4	34.0±1.4	26.5±0.4	20.9±1.6
PT_{minilm}	37.1±0.9	35.7±0.6	33.3±1.4	25.4±0.3	35.9±1.6

Notes: T1/T2/T5 contain stronger pointing cues; T3/T4 are weaker or non-pointing tiers.

TABLE III
TOP-1 ACCURACY BY REFERENCE TYPE. MEAN±STD OVER THREE RANDOM SEEDS UNDER LORO CROSS-VALIDATION.

Config	Exact NP	Pronom.	Part.
n	1,433	1,810	521
P	19.2±0.1	18.2±0.5	19.9±0.7
T_{minilm}	37.4±0.8	17.3±0.3	17.9±0.5
PT_{minilm}	45.2±0.2	23.3±0.1	25.3±0.9

Notes: Exact NP = exact noun phrase; Pronom. = pronominal reference; Part. = partitive reference.

qualitative changes in gate policy. And a single global α commits to one tradeoff across the full dataset, which (as §V-B will show) is optimal for neither the pointing nor the non-pointing regime. Both observations frame the analysis that follows.

B. Asymmetric Pose–Text Complementarity

Pose and text disagree systematically by gesture tier (Table II). On pointing references (T1∪T2∪T5, $n = 2{,}068$), pose alone achieves 27.9% top-1; text (T_{minilm}) trails at 23.2%. On non-pointing references (T3∪T4, $n = 1{,}696$), the pattern reverses: pose collapses to 7.7% while text reaches 27.2%. The per-tier view sharpens the asymmetry: on T1 (clear pointing), pose leads text by 8.7 points (31.5 vs. 22.8); on T4 (no gesture), text leads pose by 19 points (26.5 vs. 7.5). Pose stays above random on T4 (7.5% vs. $\sim 2\%$) because the head and body channels pick up non-deictic orientation cues even when arms are at rest. Neither singleton exceeds 27% on the full dataset.

An *oracle* policy that selects the best singleton per subset (pose on pointing, text on non-pointing) achieves a weighted top-1 of $\frac{2068}{3764} \cdot 27.9 + \frac{1696}{3764} \cdot 27.2 \approx 27.6\%$. Any fusion method that does not exceed this oracle is doing nothing beyond implicit subset selection.

C. Scalar Fusion Exceeds Singletons When Pathways Are Well-Specified

PT_{minilm}, fusion with frozen MiniLM category embeddings on the text pathway and learned 16-D on the pose pathway, reaches 31.9% top-1 aggregate, exceeding pose (18.8%) by 13.1 points and T_{minilm} (25.0%) by 6.9 (both significant, paired t-tests $p < 0.01$). It also exceeds the singleton-oracle (27.6%) by 4.3 points: the gate is combining complementary scores within each sample in a way that simple subset-switching cannot reproduce.

PT_{minilm} wins on every reference type (Table III). The margin is largest on exact noun phrases, where the utterance contains a matching scene-graph label: 45.2%, exceeding T_{minilm} by 7.8 points. On pronominals (23.3%) and partitives (25.3%), which carry no explicit category information, PT_{minilm} still exceeds both singletons, indicating that the benefit is not purely lexical matching. The fusion gain is visible even near pose's geometric ceiling: on T1 (clear pointing), pose alone reaches 31.5% and PT_{minilm} lifts this to 37.1%, text and pose jointly disambiguating among the pointed-at candidates in the pointing cone.

Category representation, not just category presence, shapes fusion behavior. PT with learned 16-D category embeddings reaches 28.2% aggregate, 3.7 points below PT_{minilm} (paired $t(2) = 12.4$, $p = 0.005$) despite identical architecture and training. Yet on text alone, learned (T, 22.5%) and MiniLM (T_{minilm}, 25.0%) are indistinguishable ($p = 0.323$), note the substantially higher seed variance under learned embeddings, consistent with unreliable per-category vectors at MM-Conv's training counts. The representation-choice gap *emerges specifically under fusion*, where the utterance embedding and per-object category both reside in MiniLM space for PT_{minilm}, letting the gate combine pathways more sharply. $PT_{minilm,both}$ (MiniLM on both pathways) reaches 31.6%, within noise of PT_{minilm}: the pose pathway's 6-D angular features do not benefit from a 384-D semantic augmentation.

VI. IMPLICATIONS FOR RELIABLE SEMANTIC GROUNDING

Category representation is a hidden confound in fusion evaluation. The choice of per-object category representation (learned 16-D vs. frozen MiniLM 384-D) changes fusion accuracy by 3.7 points while leaving text-alone accuracy unchanged, a dissociation only observable when category contribution is isolated per pathway. Under shared-encoder or shared-category-table architectures this effect is indistinguishable from a fusion-mechanism limitation. For semantic grounding systems whose downstream action selection depends on fusion reliability, both category representations should be reported, and pathway-local decoupling treated as a prerequisite for interpretable fusion claims.

Naturalistic referential gesture data is the right testbed. Synthetic and post-hoc benchmarks exclude the $\sim 45\%$ of references without clear gestures, eliminating the complementarity signal fusion can exploit. Fusion methods evaluated only on synthetic or post-hoc data may behave differently under naturalistic communicative pressure, where roughly half of references lack clear pointing.

Limitations and future work. Our analysis is bounded by the use of ground-truth SMPL-X; performance under estimated pose is the logical next step for deployment. While scalar fusion significantly exceeds singletons, the global gate is suboptimal for non-pointing tiers; per-sample gating conditioned on pointing-classifier output would allow $\alpha \to 0$ on non-deictic expressions. Comparisons against open-vocabulary VLM-based 3D grounding would further contextualize absolute accuracy.

REFERENCES

[1] D. Z. Chen, A. X. Chang, and M. Nießner, "Scanrefer: 3d object localization in rgb-d scans using natural language," in *European conference on computer vision*. Springer, 2020, pp. 202–221.

[2] P. Achlioptas, A. Abdelreheem, F. Xia, M. Elhoseiny, and L. Guibas, "Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes," in *European conference on computer vision*. Springer, 2020, pp. 422–440.

[3] A. M. Mane, D. Weerakoon, V. Subbaraju, S. Sen, S. E. Sarma, and A. Misra, "Ges3vig: Incorporating pointing gestures into language-based 3d visual grounding for embodied reference understanding," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 9017–9026.

[4] Y. Chen, Q. Li, D. Kong, Y. L. Kei, S.-C. Zhu, T. Gao, Y. Zhu, and S. Huang, "Yourefit: Embodied reference understanding with language and gesture," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1385–1395.

[5] A. Deichler, J. O'Regan, F. Irmak Dogan, L. Marcinek, A. Klezovich, I. Leite, and J. Beskow, "Mm-conv: A multimodal dataset and benchmark for context-aware grounding in 3d dialogue," in *Proceedings of the 15th International Conference on Language Resources and Evaluation (LREC 2026)*, 2026.

[6] A. Jain, N. Gkanatsios, I. Mediratta, and K. Fragkiadaki, "Bottom up top down detection transformers for language grounding in images and point clouds," in *European Conference on Computer Vision*. Springer, 2022, pp. 417–433.

[7] Y. Wu, X. Cheng, R. Zhang, Z. Cheng, and J. 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*, 2023, pp. 19 231–19 242.

[8] N. Hughes, Y. Chang, and L. Carlone, "Hydra: A real-time spatial perception system for 3d scene graph construction and optimization," *arXiv preprint arXiv:2201.13360*, 2022.

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

[10] Z. Wang, Y. Chen, B. Jia, P. Li, J. Zhang, J. Zhang, T. Liu, Y. Zhu, W. Liang, and S. Huang, "Move as you say interact as you can: Language-guided human motion generation with scene affordance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 433–444.

[11] A. Deichler, S. Wang, S. Alexanderson, and J. Beskow, "Learning to generate pointing gestures in situated embodied conversational agents," *Frontiers in Robotics and AI*, vol. 10, 2023.

[12] Y. Liu, D. Chi, S. Wu, Z. Zhang, Y. Zhuang, B. Yang, H. Zhu, L. Zhang, P. Xie, D. G. A. Bravo *et al.*, "Omnieva: Embodied versatile planner via task-adaptive 3d-grounded and embodiment-aware reasoning," *International Conference on Learning Representations (ICLR)*, 2026.

[13] M. M. Islam, A. Gladstone, S. Sarker, G. Nanduru, M. Fahim, K. Du, A. Chadha, and T. Iqbal, "Embodied referring expression comprehension in human-robot interaction," in *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction*, 2026, pp. 503–512.

[14] J. Yang, X. Chen, S. Qian, N. Madaan, M. Iyengar, D. F. Fouhey, and J. Chai, "Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 7694–7701.

[15] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," in *NeurIPS*, 2020.