A MULTIMODAL DEPTH-AWARE METHOD FOR EMBODIED REFERENCE UNDERSTANDING

Fevziye Irem Eyiokur[1]   *Dogucan Yaman[1]*   *Hazım Kemal Ekenel[2]*   *Alexander Waibel[1,3]*

[1]Karlsruhe Institute of Technology, [2]Istanbul Technical University, [3]Carnegie Mellon University

ABSTRACT

Embodied Reference Understanding requires identifying a target object in a visual scene based on both language instructions and pointing cues. While prior works have shown progress in open-vocabulary object detection, they often fail in ambiguous scenarios where multiple candidate objects exist in the scene. To address these challenges, we propose a novel ERU framework that jointly leverages LLM-based data augmentation, depth-map modality, and a depth-aware decision module. This design enables robust integration of linguistic and embodied cues, improving disambiguation in complex or cluttered environments. Experimental results on two datasets demonstrate that our approach significantly outperforms existing baselines, achieving more accurate and reliable referent detection.

Index Terms— Embodied reference understanding, pointing target detection, multimodal learning

1. INTRODUCTION

Embodied Reference Understanding (ERU) [1] is the task of identifying a specific object in a visual scene based on language instructions and pointing cues within the image. This task plays a key role in real-world applications such as human–robot interaction and assistive robotics where systems must determine which object a person is referring to. While open vocabulary large models [2, 3, 4] have made significant progress in detecting objects mentioned in natural language, they often fall short on the ERU task, particularly in ambiguous scenes. When multiple instances of the same object type are present, these models tend to detect all matching candidates without the ability to disambiguate and corretly identify the pointed one. Moreover, when the textual instruction itself is vague or ambiguous, these models struggle even further, often failing to identify the correct object since they generally cannot utilize pointing cue. Specifically, Large Multimodel Models (LMMs) suffer from outputting bounding boxes (bboxes) coordinates with confidence scores or sometimes even outputting incorrect format. Furthermore, fine-tuning them is prohibitively expensive, and their inference is computationally costly as well. These limitations highlight the need for additional disambiguation cues, specifically embodied gesture signals, that can resolve referential ambiguity and enable accurate identification of the intended object. Therefore, both signals are crucial for identifying the referent, aligning with early multimodal interaction studies [5, 6]. In ERU, two main challenges arise: (1) identifying candidate objects from text, as in standard grounding tasks, and (2) interpreting pointing cues from visual input to guide the final prediction. The latter requires understanding human pose, inferring pointing direction, and handling visual complexities such as perspective, occlusion, and depth.

The ERU task, which incorporates nonverbal cues (particularly pointing gestures) to resolve referential ambiguity, was first introduced in [1], where saliency and pose features were used to capture pointing direction, and was later enhanced with depth-based reasoning [7] and "virtual-touch-lines" connecting eye and fingertip [8]. More recent works adopt transformer-based multimodal detectors with heatmap-based pointing representations and extend ERU into 3D embodied settings through benchmarks such as ScanERU [9] and Ges3ViG [10]. Related areas such as Referring Expression Comprehension (REC) [11, 12, 13, 2, 14, 15, 16] and nonverbal communication studies [17, 18, 19, 20, 21] are also relevant, but often limited when relying solely on language or cues like gaze. Despite these advances, most ERU methods still overlook the explicit modeling of the pointing line and depth, leaving text as the dominant cue. Consequently, they struggle in ambiguous or cluttered scenes where pointing and depth information could provide disambiguation.

In this paper, we propose a novel approach to improve ERU by leveraging complementary training strategies and a depth-aware ensemble mechanism. First, we apply text data augmentation during training to enhance the model's generalization capacity. Second, we incorporate estimated depth maps as additional input, as depth information can provide useful spatial cues. However, we observe that combining depth with augmentation in a single model does not yield consistent improvements, since depth can both help and hinder performance depending on the scene. To address this, we train two parallel models: one with augmented text data and one with normal text (no augmentation) but with depth input. Finally, we introduce a depth-aware decision module (DADM) that integrates the predictions of the two models. Specifically, the final output is selected as the bounding box closest to the
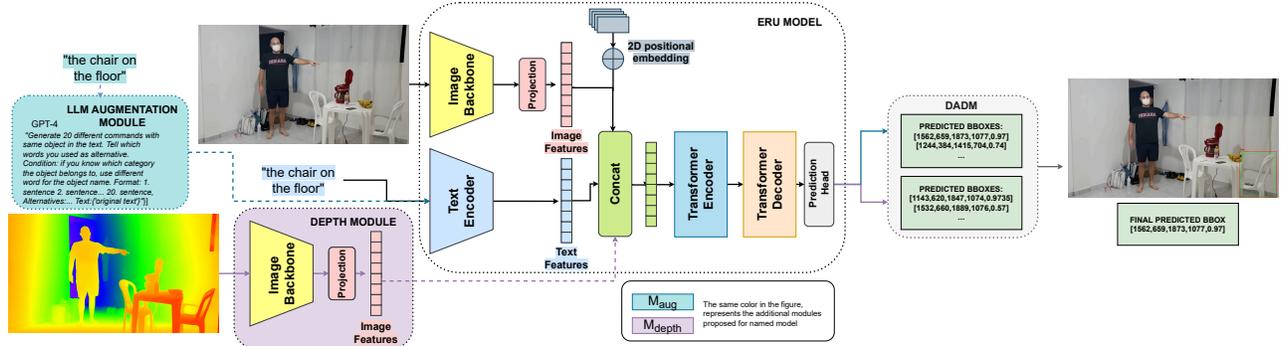
Fig. 1. Overall framework. Depth and LLM Augmentation modules are used interchangeably with proposed parallel models.

predicted pointing line out of the predicted bounding boxes from two models, enabling more robust and accurate target object detection. Our contributions are as follows. (1) We propose LLM-based text augmentation for ERU, significantly improving target detection. (2) We introduce a depth-aware ERU model to address failures caused by missing depth cues. (3) We design a depth-aware decision module that combines augmented and depth-aware models for robust predictions. (4) We achieve state-of-the-art results on two benchmarks, supported by extensive ablation studies.

2. METHODOLOGY

Text Data Augmentation. For each target object, we prompt GPT-4 [22] to generate 20 alternative sentences by replacing the object with semantically similar words only in the training set. Although the images remain unchanged, pairing each image with 20 additional sentences increases the training set from its original size to 21 times larger. This strategy improves model robustness to variation in complex text prompts.
Depth Map Estimation. We employ the Depth Pro [23] for depth map estimation. It's built on a multi-scale vision transformer, and effectively captures global context for accurate relative depth while preserving fine details for sharp boundaries. We select Depth Pro for its ability to generate seamless depth predictions directly from RGB input.

2.1. Model Architecture

In this paper, we propose two complementary models for performing ERU. The first model, M_{aug}, is trained on augmented data, whereas the second model, M_{depth}, is trained on the non-augmented data with the estimated depth map modality. Fig. 1 shows both of our models together with the final decision module. The augmented data is used exclusively in M_{aug}, while the depth module appears only in M_{depth}.
Problem Definition. Given an RGB image $x_{img} \in \mathbb{R}^{3 \times H \times W}$ and a text input $x_{text} \in \mathbb{N}^L$, the goal is to predict a bounding box $x_{bbox} \in \mathbb{R}^4$ that indicates an object referenced by the text instruction and pointing gesture.

Encoders. In both models, we use a pretrained ROBERTa encoder, a robustly optimized variant of BERT, to process the textual instruction and generate text embeddings. For the image encoder, we employ a ResNet-101 [24] to extract image features, $F_I \in \mathbb{R}^{2048 \times 8 \times 8}$. In M_{depth}, a lightweight ResNet-18 [24] encodes the depth map into features $F_{depth} \in \mathbb{R}^{256 \times 8 \times 8}$, as depth contains less information than the image and requires a smaller representation. This design choice also improves computational efficiency.
Transformer encoder-decoder. We obtain features from text, image, and depth encoders, where depth applies only to M_{depth}. Each embedding is projected to 256 channels using a 1×1 convolution, after which the spatial dimensions are flattened into token sequences. These sequences are concatenated to form the multimodal input, which is processed by a transformer encoder. The encoder output, together with a set of learnable queries, is passed to a transformer decoder that produces the final object and gesture embeddings.
Prediction head. The object and gestural embeddings generated by the transformer decoder are passed to prediction heads implemented as multi-layer perceptrons. These heads output candidate bboxes, their center points for the referent, and eye-fingertip keypoints. For each model, predictions are ranked by confidence score in descending order, and the top two are forwarded to DADM.
Training details. Overall objective function is:

$$L = \lambda_1 L_{bb} + \lambda_2 L_a + \lambda_3 L_g + \lambda_4 L_t + \lambda_5 L_c \qquad (1)$$

Here, L_{bb} is the bbox loss (L1 + GIoU), L_g represents the gestural loss, the distance between predicted and GT eye and fingertip coordinates. L_t and L_c are soft token and contrastive losses, respectively, following [25]. The alignment loss, L_a, penalizes misalignment between the eye–fingertip and the eye–object vectors by comparing their cosine similarity to the ground truth, encouraging predictions aligned with pointing. We found best coefficients respectively $= (2, 1, 10, 1, 1)$.
Pointing Line Estimation. Prior work [8] shows that the referent object typically lies along the line connecting a person's head and pointing finger. Following this intuition, we use the

Algorithm 1 Depth-Aware Decision Module (DADM)

Require: Bounding box of Top-2 predictions from M_{aug} and M_{depth}: $B_{aug} = [(b^0_{aug}, c^0_{aug}), (b^1_{aug}, c^1_{aug})]$, $B_{depth} = [(b^0_{depth}, c^0_{depth}), (b^1_{depth}, c^1_{depth})]$
Require: Thresholds: $T_1 = 0.9$(IoU), $T_2 = 0.6$ (confidence)
Require: Predicted pointing line map I_L
Ensure: Final prediction b^*
1: Compute IoU: $iou \leftarrow$ IoU(b^0_{aug}, b^0_{depth})
2: **if** $iou \geq T_1$ **then**
3:     $b^* \leftarrow b^0_{aug}$
4: **else**
5:     Initialize candidate list: Candidates $\leftarrow \{b^0_{aug}, b^0_{depth}\}$
6:     **for** each $(b^1, c^1) \in \{(b^1_{aug}, c^1_{aug}), (b^1_{depth}, c^1_{depth})\}$ **do**
7:         **if** $c^1 \geq T_2$ **then**
8:             Add b^1 to Candidates
9:         **end if**
10:     **end for**
11:     Compute distance of each candidate to I_L
12:     $b^* \leftarrow$ candidate with shortest distance to I_L
13: **end if**
14: **return** b^*

OpenPose [26] pose estimation model to detect approximate eye and fingertip coordinates, and draw a conic line from the detected eye to the fingertip. This pointing line is then used in DADM to calculate distance for the final prediction.

2.2. Depth-Aware Decision Module

We finally introduce a novel Depth-Aware Decision Module (DADM), see Algorithm 1, which incorporates the top two predictions from M_{aug} and M_{depth} to determine the final output. The module first computes the intersection-over-union (IoU) between the highest-confidence bbox predictions of M_{aug} and M_{depth}. If their overlap exceeds a threshold T_1, the final prediction is set to b^0_{aug}, since M_{aug} generally demonstrates higher accuracy. If the overlap is below T_1, we instead use the predicted pointing line map and select the bounding box with the shortest distance to the line. Because the top predictions from both models may fail to overlap, especially in complex or ambiguous scenes, we also consider the second-ranked predictions. To avoid low-quality candidates, we include these only if their confidence score exceeds a second threshold T_2. A candidate list is then formed from the valid predictions, and the bbox whose center lies closest to the pointing line is chosen as the final output.

3. EXPERIMENTAL RESULTS

Datasets. For training, we utilize the YouRefIt [1], the widely used benchmark. We evaluate our models on the YouRefIt test set and the unseen ISL pointing dataset [20].

Evaluation. For evaluation metrics and setup, we follow prior work [1]. IoU is computed at three threshold values: 0.25, 0.50, and 0.75. Additionally, objects are categorized as *small (S)*, *medium (M)*, and *large (L)* based on their bounding box size, and scores are reported with respect to object size.

Comparison. We present quantitative results in Table 1. The first section of the table reports results from recent LMMs [3, 4] and the SOTA open-set object detector Grounding DINO [2]. We evaluate whether these models can perform the task without fine-tuning, leveraging their zero-shot capabilities. The second section summarizes results from recent prior works on this task. The third section presents the performance of our models. Here, *baseline* refers to straightforward training of our architecture with YouRefIt dataset; *Only Aug* corresponds to M_{aug} trained with text augmented data; *Only Depth* shows the performance of M_{depth}; and *Full* represents the combination of M_{aug} and M_{depth} with the DADM.

Results. Our baseline model achieves competitive performance compared to the previous model [8]. The M_{aug} model clearly outperforms it at the 0.25 and 0.50 IoU thresholds. Although M_{depth} is less accurate than the augmented model overall, analysis of its outputs reveals that in critical cases, where most models fail due to missing depth information, it performs robustly thanks to depth awareness. This motivates combining M_{depth} with M_{aug}. However, trainin M_{depth} with augmented data harms the performance. Therefore, we introduce DADM, we obtain a substantial performance boost and achieve SOTA performance with *Full* model, except for *small* and *medium* objects at 0.75 threshold. The evaluated LMMs perform significantly worse than our proposed method. We further evaluate our full model on the unseen and more challenging ISL pointing dataset [20], comparing its performance against the previous best perfomed models. The results show that our model surpasses both Touch-Line models and LMMs by a large margin at $IoU = 0.25$ and $IoU = 0.50$ thresholds, while achieving comparable performance at $IoU = 0.75$ with LMMs and Touch-line-VTL. Note that the 0.75 threshold primarily reflects the precision of predicted bounding box coordinates. Our model is more accurate in identifying the correct target object in the scene, though it occasionally produces slightly less precise bounding boxes. In Fig. 2, we present qualitative results together with the predicted depth maps.

Ablation Study for Decision Method in DADM. The first part of Table 3 presents an ablation study of different decision strategies within DADM. A selects the top-1 predictions from both models and chooses the box with the shortest distance to the pointing line. B, instead, selects the box with the largest overlap in pixels with the pointing line. Finally, C normalizes this overlap by bounding-box size and selects the box with the highest percentage. While all methods yield performance gains, our proposed DADM achieves the most significant improvement.

Ablation Study for Final Decision. The second part of Table 3 compares alternative strategies to DADM for selecting the
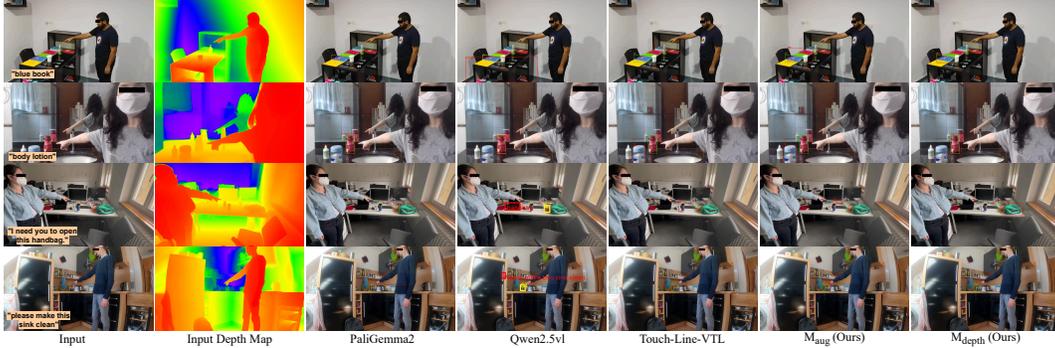
Fig. 2. Visual comparisons on YouRefIt dataset (first two rows) and ISL pointing dataset (last two rows).

IoU Threshold for mAP	0.25				0.50				0.75			
Object Sizes	All	S	M	L	All	S	M	L	All	S	M	L
PaliGemma2 [3]	58.8	29.0	53.5	75.8	46.9	22.1	50.8	68.0	31.7	6.2	34.1	54.8
Qwen2.5vl [4]	38.9	17.0	41.8	58.0	31.0	11.1	33.6	48.1	20.0	5.7	19.8	34.5
Grounding DINO [2]	57.9	38.0	60.9	74.9	54.9	35.7	59.3	69.6	**42.3**	**22.7**	**45.9**	_58.4_
FAOA [27]	44.5	30.6	48.6	54.1	30.4	15.8	36.5	39.3	8.5	1.4	9.6	14.4
ReSC [28]	49.2	32.3	54.7	60.1	34.9	14.1	42.5	47.7	10.5	0.2	10.6	20.1
YourRefit PAF [1]	52.6	35.9	60.5	61.4	37.6	14.6	49.1	49.1	12.7	1.0	16.5	20.5
YourRefit Full [1]	54.7	38.5	64.1	61.6	40.5	16.3	54.4	51.1	14.0	1.2	17.2	23.3
REP [7]	58.8	44.7	68.9	63.2	45.7	25.4	57.7	54.3	18.8	3.8	22.2	29.9
Touch-Line-EWL [8]	69.5	56.6	71.7	80.0	60.7	44.4	66.2	71.2	35.5	11.8	38.9	55.0
Touch-Line-VTL [8]	71.1	55.9	75.5	81.7	63.5	47.0	_70.2_	73.1	_39.0_	_13.4_	_45.2_	57.8
Baseline	71.2	59.5	73.0	80.8	60.1	43.2	66.4	70.5	32.8	8.6	35.4	53.4
M_{aug} (Only Aug)	_73.2_	_60.7_	_75.7_	_83.5_	_64.1_	_47.5_	69.8	_75.3_	35.2	12.7	37.2	55.8
M_{depth} (Only Depth)	70.8	56.5	73.3	82.7	60.8	43.5	67.9	71.4	33.0	9.5	37.7	52.1
M_{aug_depth} (Aug + Depth)	66.1	56.1	69.7	72.4	53.6	37.9	61.7	61.2	23.3	6.5	22.0	40.7
DA-ERU (Full)	**78.7**	**65.7**	**82.1**	**88.4**	**67.6**	**48.5**	**75.4**	**79.3**	38.1	_13.4_	39.8	**61.0**

Table 1. Comparison of our model with prior work in terms of mean Average Precision (mAP) at different IoU thresholds, across various object sizes, on the YouRefIt dataset [1].

Setup	IoU=0.25	IoU=0.50	IoU=0.75
PaliGemma2 [3]	47.2	39.5	**31.6**
Qwen2.5vl [4]	32.5	32.1	29.2
Touch-Line-EWL [8]	45.0	35.8	22.0
Touch-Line-VTL [8]	47.7	36.7	17.4
DA-ERU (Full)	**62.7**	**50.1**	30.1

Table 2. Test results on unseen ISL pointing dataset [20].

Setup	IoU=0.25	IoU=0.50	IoU=0.75
A - Distance to Pointing Line	75.5	64.8	35.8
B - Overlapping area w/ Pointing Line	75.9	64.7	35.9
C - Overlapping area w/ Pointing Line Percentage	75.1	64.4	35.9
Confidence score based	73.3	63.8	36.3
Adaptive depth	73.3	63.0	35.9
DADM (Distance)	**78.7**	**67.6**	**38.1**

Table 3. Ablation study for DADM and distance strategy.

final object. The first confidence score based method selects between the top-1 predictions of both models by choosing the box with the higher confidence score. Adaptive depth also uses the top-1 predictions, but selects the model whose confidence gap between the first and second prediction is larger, indicating a clearer decision. Among these strategies, DADM consistently achieves the best performance.

4. CONCLUSION

We address the challenges of ERU by tackling the limitations of existing methods. Our contributions, text augmentation, depth estimation, and a depth-aware decision module, enhance pointing target detection. Experiments show that both text augmentation and incorporating depth maps improve performance individually. More importantly, combining these models with our novel depth-aware decision module (DADM), which leverages the predicted pointing line for distance-based selection, yields more accurate and robust referent understanding in complex and ambiguous visual scenes.

5. REFERENCES

[1] Yixin Chen, Qing Li, Deqian Kong, Yik Lun Kei, Song-Chun Zhu, Tao Gao, Yixin Zhu, and Siyuan Huang, "Yourefit: Embodied reference understanding with language and gesture," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1385–1395.

[2] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al., "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," in *ECCV*. Springer, 2024, pp. 38–55.

[3] Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al., "Paligemma 2: A family of versatile vlms for transfer," *arXiv preprint arXiv:2412.03555*, 2024.

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al., "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[5] Jie Yang, Rainer Stiefelhagen, Uwe Meier, and Alex Waibel, "Visual tracking for multimodal human computer interaction," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1998, pp. 140–147.

[6] Bernhard Suhm, Brad Myers, and Alex Waibel, "Model-based and empirical evaluation of multimodal interactive error correction," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 1999, pp. 584–591.

[7] Cheng Shi and Sibei Yang, "Spatial and visual perspective-taking via view rotation and relation reasoning for embodied reference understanding," in *European Conference on Computer Vision*. Springer, 2022, pp. 201–218.

[8] Yang Li, Xiaoxue Chen, Hao Zhao, Jiangtao Gong, Guyue Zhou, Federico Rossano, and Yixin Zhu, "Understanding embodied reference with touch-line transformer.," in *International Conference on Learning Representations*, 2023.

[9] Ziyang Lu, Yunqiang Pei, Guoqing Wang, Peiwei Li, Yang Yang, Yinjie Lei, and Heng Tao Shen, "Scaneru: Interactive 3d visual grounding based on embodied reference understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 3936–3944.

[10] Atharv Mahesh Mane, Dulanga Weerakoon, Vigneshwaran Subbaraju, Sougata Sen, Sanjay E Sarma, and Archan 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.

[11] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg, "Referitgame: Referring to objects in photographs of natural scenes," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 787–798.

[12] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg, "Mattnet: Modular attention network for referring expression comprehension," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1307–1315.

[13] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao, "Clip-adapter: Better vision-language models with feature adapters," *IJCV*, vol. 132, no. 2, pp. 581–595, 2024.

[14] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui, "Open-vocabulary object detection via vision and language knowledge distillation," *arXiv preprint arXiv:2104.13921*, 2021.

[15] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang, "Open-vocabulary object detection using captions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14393–14402.

[16] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al., "Paligemma: A versatile 3b vlm for transfer," *arXiv preprint arXiv:2407.07726*, 2024.

[17] Rainer Stiefelhagen, Michael Finke, Jie Yang, and Alex Waibel, "From gaze to focus of attention," in *International Conference on Advances in Visual Information Systems*. Springer, 1999, pp. 765–772.

[18] Rainer Stiefelhagen, Christian Fugen, R Gieselmann, Hartwig Holzapfel, Kai Nickel, and Alex Waibel, "Natural human-robot interaction using speech, head pose and gestures," in *2004 IEEE/RSJ IROS*. IEEE, 2004, vol. 3, pp. 2422–2427.

[19] Akira Oyama, Shoichi Hasegawa, Hikaru Nakagawa, Akira Taniguchi, Yoshinobu Hagiwara, and Tadahiro Taniguchi, "Exophora resolution of linguistic instructions with a demonstrative based on real-world multimodal information," in *IEEE International Conference on Robot and Human Interactive Communication*. IEEE, 2023, pp. 2617–2623.

[20] Stefan Constantin, Fevziye Irem Eyiokur, Dogucan Yaman, Leonard Bärmann, and Alex Waibel, "Interactive multimodal robot dialog using pointing gesture recognition," in *European conference on computer vision*. Springer, 2022, pp. 640–657.

[21] Stefan Constantin, Fevziye Irem Eyiokur, Dogucan Yaman, Leonard Bärmann, and Alex Waibel, "Multimodal error correction with natural language and pointing gestures," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1976–1986.

[22] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al., "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[23] Aleksei Bochkovskii, AmaÃĞl Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun, "Depth pro: Sharp monocular metric depth in less than a second," *arXiv preprint arXiv:2410.02073*, 2024.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[25] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion, "Mdetr-modulated detection for end-to-end multi-modal understanding," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1780–1790.

[26] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[27] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo, "A fast and accurate one-stage approach to visual grounding," in *ICCV*, 2019, pp. 4683–4693.

[28] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo, "Improving one-stage visual grounding by recursive sub-query construction," in *ECCV*. Springer, 2020, pp. 387–404.