# Look and Tell: A Dataset for Multimodal Grounding Across Egocentric and Exocentric Views

**Anna Deichler**
TMH
KTH Royal Institute of Technology
deichler@kth.se

**Jonas Beskow**
TMH
KTH Royal Institute of Technology
beskow@kth.se

## Abstract

We introduce Look and Tell, a multimodal dataset for studying referential communication across egocentric and exocentric perspectives. Using Meta Project Aria smart glasses and stationary cameras, we recorded synchronized gaze, speech, and video as 25 participants instructed a partner to identify ingredients in a kitchen. Combined with 3D scene reconstructions, this setup provides a benchmark for evaluating how different spatial representations (2D vs. 3D; ego vs. exo) affect multimodal grounding. The dataset contains 3.67 hours of recordings, including 2,707 richly annotated referential expressions, and is designed to advance the development of embodied agents that can understand and engage in situated dialogue.

## 1 Introduction

Understanding how humans coordinate gaze, gestures, and speech in referential communication is critical for building embodied agents that can interact naturally and exhibit spatial intelligence. While gaze has long been recognized as a cue for attentional focus, most research has emphasized its temporal relationship to speech in controlled tasks. Less attention has been given to how spatial representations of the environment—whether in 2D image space, 3D reconstructions, or across egocentric and exocentric perspectives—influence multimodal spatial grounding.

We present a new dataset collected with 25 participants in a naturalistic kitchen setting, where participants recalled recipe ingredients while wearing Meta Aria smart glasses [Engel et al., 2023]. The glasses provided synchronized gaze and speech streams, complemented by exocentric GoPro recordings. This dual-view setup enables both fine-grained analysis of gaze–speech synchrony and systematic evaluation of representation choices in multimodal grounding. Specifically, the dataset supports comparisons between 2D and 3D scene representations, as well as between egocentric and exocentric perspectives, offering a unique testbed for investigating how humans and embodied models organize spatial knowledge during communication. This is crucial for human–robot interaction, where an agent must understand references from a user's point of view and integrate multimodal signals such as gaze, gestures, and speech. This dual-view paradigm is critical for tackling challenges in **shared autonomy and human-robot collaboration**, where an agent must simultaneously interpret a user's first-person intent while maintaining an objective, third-person model of the world state.

Our contributions are threefold: (1) a multimodal dataset of synchronized gaze, speech, and dual-view video designed to study **situated dialogue**; (2) an annotation pipeline for **aligning spatial concepts across language and vision**; and (3) a new **benchmark for evaluating spatial intelligence** in **grounded communication**. By capturing rich, natural behavior, our dataset provides a foundation for future analysis of how gesture, gaze, and speech jointly enable **multimodal spatial grounding**.

## 2 Related Work

**Gaze and speech in communication.** Prior work shows that gaze strongly predicts communicative goals [Hanna and Brennan, 2007, Brennan and Hanna, 2007], and has been combined with gestures or speech to improve reference resolution [Renner et al., 2014]. Event synchronization studies [Kaur et al., 2003] highlight the potential of aligning multimodal streams, though mainly in controlled tasks. Recent work emphasizes gaze in instruction and teaching contexts [Wagner et al., 2023] and in virtual environments [Tanriverdi and Jacob, 2000]. However, temporal alignment of gaze and speech in natural referential tasks remains underexplored.

**Multimodal communication datasets.** Several datasets integrate multiple communicative channels, such as speech, gaze, and gesture in collaborative tasks [Kontogiorgos et al., 2018]. VENUS [Kim et al., 2025] combines speech, facial expressions, and body pose. In vision–language research, gaze has been used for captioning [Sugano and Bulling, 2016], VQA [Vasudevan et al., 2018, Ilaslan et al., 2023]. Despite this progress, datasets combining first-person gaze with concurrent speech in referential tasks are rare.

**Spatial representations for grounding.** Recent research highlights the importance of spatial representations for multimodal agents. Ego4D and related egocentric datasets emphasize large-scale video and gaze [Grauman et al., 2022], while 3D QA datasets such as ScanQA [Gao et al., 2021] and EmbodiedQA [Das et al., 2018] explore grounding in reconstructed scenes. Yet, few datasets allow direct comparison between 2D vs. 3D representations or ego vs. exo perspectives in the same task. Our dataset bridges this gap, enabling systematic evaluation of how representational choices affect grounding in situated referential communication.

## 3 Data Collection & Dataset Description

### 3.1 Experiment Design

**Participants.** 25 participants (18 female, 7 male; age range 22–37) were recruited at KTH Kitchen Lab. **Materials.** Aria smart glasses (for eye tracking and audio recording), GoPro cameras (environment capture), recipes, and food ingredients. **Task.** Each participant memorized a step of a recipe and then instructed it to a conversational partner. While speaking, they were asked to identify and refer to the corresponding ingredients in the environment.

**Procedure.**

1. Participant reads recipe step.
2. Participant recalls step aloud while locating ingredients.
3. Egocentric Aria glasses record synchronized gaze, audio, video, and pointcloud data; exocentric GoPro records audio and video feed.
4. Sessions were recorded across five recipes per participant.

### 3.2 Dataset Description

The **Look and Tell** dataset, curated by KTH Royal Institute of Technology, offers a unique resource for investigating the interplay between visual attention and spoken language. This dataset was collected using Aria smart glasses, which allowed for the simultaneous and real-time capture of participants' eye-tracking data and speech audio.

The experimental paradigm involved participants identifying various food items from a recipe. This task was designed to elicit natural referential communication, providing rich, ecologically valid data on how individuals visually fixate on objects while verbally describing or identifying them. The recordings from the Aria glasses provide a synchronized stream of high-resolution eye-tracking data and corresponding speech, enabling detailed analysis of gaze-speech synchronization patterns. This is complemented by exocentric GoPro recordings. This dual-view setup enables fine-grained analysis of gaze–speech synchrony and evaluation of representational choices in multimodal grounding (see Fig. 1).

Table 1: KTH-ARIA Referential Dataset Summary Statistics

| | |
|---|---|
| **Participants / Sessions** | 25 participants / 125 unique sessions |
| **Total RGB Frames** | 396,208 frames |
| **Total Duration** | 3.67 h (220.1 min) at 30 fps |
| **Frames per Session** | Mean: 3,169.7 $\pm$ 1,328.9<br>Range: 1,223–7,730 |
| **Duration per Session** | Mean: 1.8 min $\pm$ 0.7 min<br>Range: 0.7 min–4.3 min |

### 3.3 Data Modalities and Format

- **Eye-tracking data:** fixation events, gaze vectors.
- **Speech audio:** raw speech signal, later transcribed and segmented with WhisperX.
- **Video recordings:**
  - Egocentric view: raw video from the head-mounted Aria cameras (MP4 format), recorded at up to $1408 \times 1408$ pixels and 30 fps, synchronized with audio and gaze.
  - Exocentric view: side-view video recordings from stationary GoPro cameras, providing third-person context of the participant and environment.
- **Pointcloud representation** of the scene.
- **Recipe and ingredient metadata**, including surface positions.

**3D Room Reconstruction.** In addition to synchronized gaze and speech recordings, we also obtained reconstructed 3D models of the kitchen environment from separate video sessions. Using the Meta Project Aria MPS service [Meta Reality Labs Research, 2023], we extracted image frames from room recordings and generated point clouds that were then canonicalized into a shared coordinate system (Fig. 2). This process aligns multiple reconstructions from different recording sessions, yielding a consistent 3D representation of the environment. These reconstructions provide additional spatial context beyond 2D video, enabling research on how embodied agents can exploit 3D scene structure for referential grounding and multimodal interaction.

### 3.4 Dataset Statistics

We recruited 25 participants (aged 22–37; 18 female, 7 male). Across 125 sessions, the dataset contains a total of **396,208 RGB frames** ($\sim$3.67 hours at 30 fps). Sessions lasted on average $1.8 \pm 0.7$ minutes (range 0.7–4.3), corresponding to $3{,}170 \pm 1{,}329$ frames per session (range 1,223–7,730). The longest session was 4.3 minutes and the shortest 0.7 minutes. See details on 1.

Participants reported the following native languages: Korean (1), Indian languages (2), Icelandic (1), Chinese Mandarin (10), Spanish (2), Swedish/English bilingual (1), Swedish (4), Turkish (1), Nigerian (1), Indonesian (1), Portuguese (1), and Russian (1).

## 4 Annotation Pipeline and Analysis

To enable fine-grained analysis of gaze–speech synchrony, we developed a multi-stage annotation pipeline that integrates audio transcription, language-based reference extraction, and multimodal object detection.

**Audio acquisition.** Audio data was obtained from the **Meta Project Aria MPS service**, which allowed us to request synchronized audio streams from the original `.vrs` files alongside the corresponding image frames. The images were pre-processed to remove fisheye distortion, yielding undistorted RGB frames aligned with the audio.

**Speech transcription.** We transcribed the audio using **WhisperX** [Bain et al., 2023], which provides robust automatic speech recognition together with **word-level timestamps**. These timestamps form the backbone of our temporal alignment between speech and other modalities.

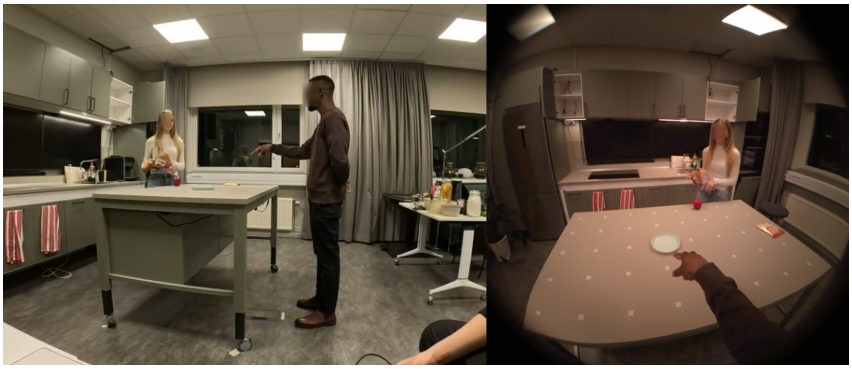


Figure 1: Example of synchronized video feeds. The exocentric (left) view provides situational context, while the egocentric (right) view captures the participant's first-person perspective, including their gaze target (green circle, overlaid for visualization).

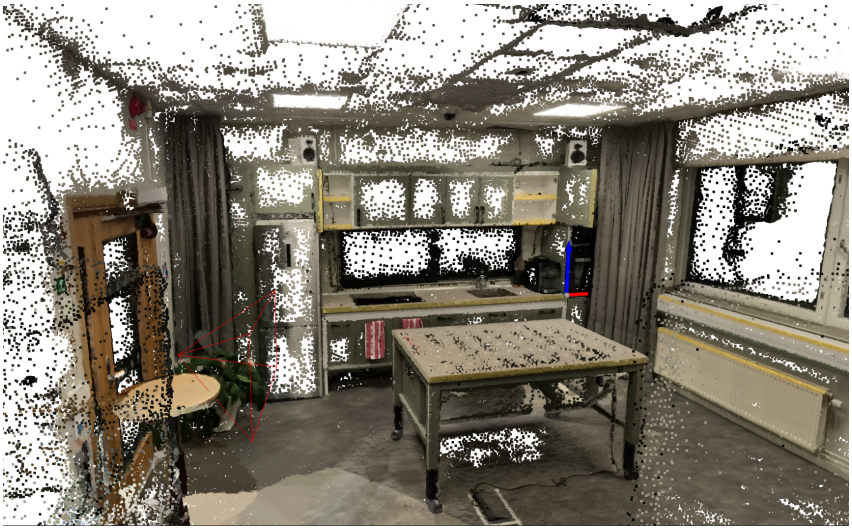


Figure 2: Example of reconstructed 3D room point cloud used as the canonical reference space. All point clouds extracted from Aria recordings are aligned and canonicalized to this shared coordinate system.

**Reference extraction.** Using the transcribed text, we employed **GPT-based prompting** to identify mentions of ingredients and distractor objects referenced in the recipe instructions, as well as additional objects spontaneously mentioned by participants [OpenAI, 2023]. Here, we use *additional object* to denote non-ingredient items from the recipe/distractor lists that participants sometimes referred to explicitly (e.g., *fridge*, *sink*, *stove*). This step produced a set of candidate **mentions** linked to precise word spans in the transcripts. The full prompting templates used for this step are provided in the Appendix.

### 4.1 GPT-based Mention Linking

To process raw speech transcripts into structured data for multimodal analysis, we automatically convert WhisperX word-level transcripts into span-level ingredient, utensil, and object mentions using a hybrid pipeline (Fig. 3). This process leverages a large language model augmented with deterministic rules to ensure consistency.

**Inputs.** The pipeline uses three sources of information: (a) tokenized transcripts with word-level timings from WhisperX, (b) recipe metadata and ingredient lists from `recipes.json`, and (c) a pre-defined set of scene distractors (which are ingredients not used in current recipe).

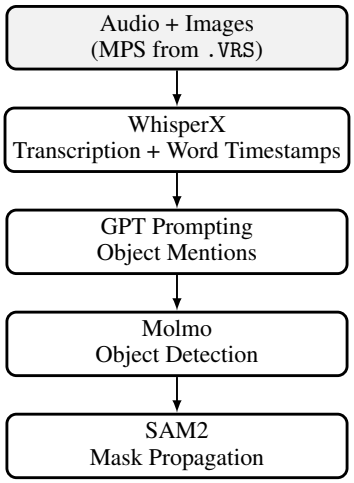



Figure 3: Annotation pipeline for the egocentric camera: synchronized audio and images are processed with WhisperX for word-level transcripts, GPT for object mentions, Molmo for object detection, and SAM2 for mask propagation.

**LLM Labelling.** We prompt a GPT model with the lower-cased transcripts and a list of candidate items (ingredients + distractors), constraining its output to a JSON schema. The model identifies and labels mention spans, providing attributes such as match type, confidence, and optional coreference links. Coreference resolution for pronouns and demonstratives (*it, this, that, them*) is guided by nearby cooking-related actions.

**Post-processing.** To ensure accuracy and consistency, we apply a series of deterministic rules: (1) normalization of predicted strings to canonical recipe or distractor names; (2) alias enrichment to handle synonyms (e.g., *tap* $\rightarrow$ *sink*); (3) precise alignment of timings and surface text; and (4) construction of a mention graph with unique IDs and antecedent links for coreferences. Automated checks are used to flag unmapped items, missing antecedents, or timestamp errors.

Full details on the prompt schema, processing rules, and handling of edge cases are documented in Appx. A.1–A.

### 4.2 Object detection and tracking.

For each candidate mention, we applied **Molmo** Deitke et al. [2024], a multimodal vision–language model, to localize the referenced item (ingredient or object; see definition above) in the undistorted frames. We prompted Molmo with concise imperatives (e.g., *"Point to the {description} object."*) and evaluated it on frames sampled from the mention interval, using at least $\max(n_{\text{frames}}, n_{\text{min}})$ frames. Molmo returned 2D point coordinates (in percent of image width/height), which we converted into seed locations to initialize the **SAM2**Ravi et al. [2024] tracker. SAM2 then propagated segmentation masks across the full mention interval, yielding dense per-frame masks aligned to each mention. When Molmo did not yield a reliable localization, we performed *manual annotation* to seed or correct the tracker, particularly for tiny items (e.g., spice containers) and visually similar look-alikes (e.g., sugar vs. wheat-flour jars; see Fig. 4). In total, 747 mentions required manual handling: 106 cases were skipped because the object was not visible in the frames, while 641 mentions were manually annotated after inspection identified them as either too small (Molmo never reliable) or belonging to ambiguous object categories.

Together, the pipeline yields synchronized **speech transcripts**, **object mentions** (tokens $\rightarrow$ mentions $\rightarrow$ chains), **frame-level points**, **bounding boxes**, and **per-frame masks** for each referential episode.

### 4.3 Analysis of Mention Annotations

Across all 25 participants and 125 sessions, we annotated **2,707 mentions** ($\sim$22 per session). Table 2 shows the distribution: ingredients dominate (62%), while **23% are pronouns or coreferential**

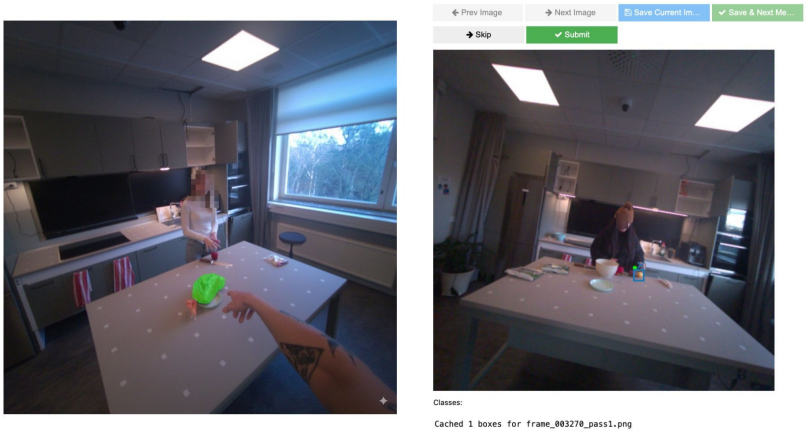



Figure 4: Examples of annotation strategies: (a) Molmo detection based automated SAM2 mask propagation, (b)manual seeding of tracker with annotation interface.

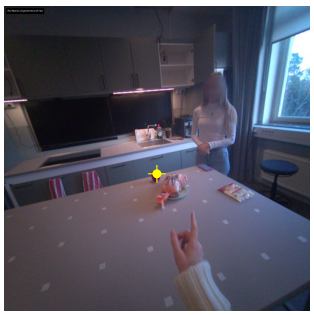
| **Mention type** | **Count** | **Proportion** |
|---|---|---|
| Ingredients | 1,680 | 62.1% |
| Pronoun/coref | 614 | 22.7% |
| Additional objects | 154 | 5.7% |
| Distractors | 28 | 1.0% |
| **Total** | 2,707 | 100% |

Table 2: Distribution of annotated mention types across the dataset.

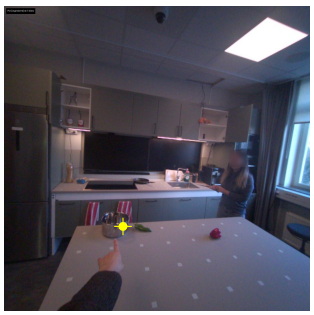
**forms**, highlighting the prevalence of indirect reference. Objects (6%) and distractors ($< 2\%$) provide additional ecological variation.

Coreference chains averaged **6.7 mentions**, indicating repeated reference to the same item once introduced. Ingredient coverage was high: nearly all target items appeared at least once (mean coverage $> 90\%$).

Although modest in scale, these annotations capture diverse referential behavior, including explicit naming, pronouns, long chains, and variable timing—features crucial for studying multimodal reference resolution. Examples of annotated mentions are shown in Fig. 5.



Paprika mention
(“slice the paprika”)



Pot mention
(“put some water in it”)

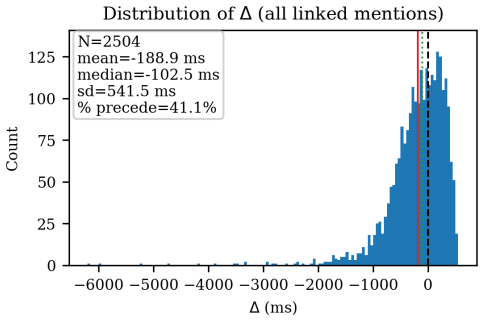
Knife mention
(“grab the knife”)

Figure 5: Examples of annotated mentions: (left) paprika, (middle) pot, (right) knife.

### 4.4 Gaze–Speech Synchrony Analysis

To quantify the temporal relationship between gaze and speech, we linked 2,504 mentions (92.5% of all 2,707 mentions) to at least one fixation event. We define $\Delta$ as fixation offset minus mention onset. Figure 6 (a) shows the distribution of the time lag $\Delta$ between mention onsets and fixation offsets, where $\Delta > 0$ indicates gaze precedes speech. The mean lag was $-189$ ms (median $-102$ ms, SD $= 541$ ms), with values ranging from $-6159$ ms to $+531$ ms. Confidence intervals around the mean were $[-211, -168]$ ms. In total, 41.1% of mentions were preceded by gaze (95% CI [39.2%, 43.1%]). The average temporal overlap between gaze and mention was 352 ms (median 367 ms).

These results indicate that while gaze and speech often overlap, gaze preceded the verbal mention in 41.1% of instances. In this naturalistic task, this suggests that speakers frequently initiate an utterance while still fixating on the referent, rather than consistently shifting their gaze before speaking.




(a) Distribution of gaze–speech lag $\Delta$ (fixation offset relative to mention onset). Positive values indicate gaze precedes speech.

| Metric | Value | 95% CI |
|---|---|---|
| Total mentions | 2707 | – |
| Linked mentions | 2504 (92.5%) | – |
| $\Delta$ mean (ms) | $-188.9$ | $[-210.7, -167.9]$ |
| $\Delta$ median (ms) | $-102.5$ | – |
| $\Delta$ SD (ms) | 541.5 | – |
| $\Delta$ range (ms) | $[-6159, +531]$ | – |
| % gaze precedes | 41.1% | [39.2%, 43.1%] |
| Overlap mean (ms) | 351.9 | – |
| Overlap median (ms) | 366.6 | – |

(b) Summary statistics for gaze–speech synchrony.

Figure 6: Gaze–speech synchrony: (a) distribution of lag $\Delta$ between mention onsets and fixation offsets, and (b) summary statistics.

## 5 Dataset Card and Ethical Considerations

**Dataset Maintenance.** The Look and Tell dataset, including all annotations and supplementary materials, will be made publicly available upon publication. It will be hosted on Hugging Face Datasets [Lhoest et al., 2021] to ensure long-term accessibility. We commit to maintaining the dataset, addressing issues raised by the community, and providing a clear versioning system for any future updates, such as the gesture annotations mentioned in our future work. A website with documentation and tutorials will also be provided at: `https://huggingface.co/datasets/annadeichler/KTH-ARIA-referential`.

**Ethical Considerations.** Participants provided informed written consent for the recording of their video, audio, and gaze, and for the use of this data in anonymized form for research purposes.

**Participant Privacy.** To protect participant privacy, all faces in the egocentric and exocentric video streams have been blurred using a standard face detection and blurring algorithm. While voices are not altered to preserve the speech data, we have reviewed the transcripts to ensure no personally identifiable information beyond what is inherent in the task was mentioned.

**Limitations and Bias.** Our dataset, while valuable, has limitations. The data was collected in a single kitchen environment (KTH Kitchen Lab), which may introduce environmental biases. The participant pool of 25 individuals, while diverse in native language, is primarily composed of university students and staff (aged 22-37), and may not be representative of the broader population in terms of age or background. Furthermore, the tasks are centered around specific recipes, which may favor individuals with some cooking familiarity. These factors should be considered when generalizing findings from this dataset.

## 6 Discussion

A core challenge in developing embodied AI is achieving robust multimodal spatial grounding, the ability to connect language to objects and locations within a shared 3D environment. Our dataset provides a unique opportunity to study how gaze, gestures, and speech unfold together in situated referential communication to achieve this. By uniquely combining egocentric (first-person) and exocentric (third-person) recordings with metadata supporting both 2D and 3D scene representations, our dataset offers a controlled **benchmark for evaluating spatial intelligence**. It is explicitly designed to facilitate research into **multimodal spatial grounding**, supporting analyses of how gaze can act as a predictive cue for resolving referential expressions and how different spatial representations may influence grounding performance. We hope this resource will foster the cross-disciplinary dialogue necessary to build embodied agents that can truly understand and communicate about space.

**Future Work.** The recorded sessions contain rich gestural behavior, including pointing and co-speech gestures, which are essential for multimodal reference resolution. Future extensions of this dataset will therefore include systematic gesture annotation, enabling analysis of how gaze, gesture, and speech jointly contribute to referential communication. This will allow embodied agents to be benchmarked not only on gaze-speech synchrony but also on their ability to integrate gestural cues into grounding and dialogue.

## Acknowledgments

We thank Kristín Hafsteinsdóttir and Yu Lu for their assistance with data collection and annotation.

## References

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*, 2023. URL `https://arxiv.org/abs/2303.00747`.

Susan E Brennan and Joy E Hanna. Coordinating attention in dialogue. *Discourse Processes*, 43(1): 51–77, 2007.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv e-prints*, pages arXiv–2409, 2024.

Jakob Engel, Thomas Whelan, Richard Newcombe, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. URL `https://arxiv.org/abs/2308.13561`.

Chen Gao, Jiarui Liu, Xin Wang, Peng Zhao, Yujing Shen, Jiajun Zhu, Yanwei Fu, and Qi Liu. Scanqa: 3d question answering with spatial reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Kristen Grauman, Alex Westbury, Eugene Zhang, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Joy E Hanna and Susan E Brennan. Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, 57(4):596–615, 2007.

M. Furkan Ilaslan et al. Gazevqa: A video question answering dataset for collaborative tasks with eye-gaze information. In *EMNLP*, 2023.

Jasleen Kaur, Marc Eaddy, and W Keith Edwards. Synchronization of gaze, speech, and gesture in multimodal human-computer interaction. In *Proceedings of the 5th International Conference on Multimodal Interfaces*, pages 1–8. ACM, 2003.

Youngmin Kim, Jiwan Chung, Jisoo Kim, Sunghyun Lee, Sangkyu Lee, Junhyeok Kim, Cheoljong Yang, and Youngjae Yu. Speaking beyond language: A large-scale multimodal dataset for learning nonverbal cues from video-grounded dialogues. In *ACL (Long Papers)*, 2025.

Dimitris Kontogiorgos, Jörg Bergmann, Giorgio Caldarola, and Stefan Kopp. A multimodal corpus for mutual gaze and joint attention in multiparty situated interaction. In *Proceedings of the LREC Workshop on Multimodal Corpora of Communication (MMC)*, 2018.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, and et al. Datasets: A community library for natural language processing. In *EMNLP: System Demonstrations*, 2021. URL `https://arxiv.org/abs/2109.02846`.

Meta Reality Labs Research. Project aria tools. `https://github.com/facebookresearch/projectaria_tools`, 2023. See citation guidance in docs.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Patrick Renner, Thies Pfeiffer, Nadine Pfeiffer-Leßmann, and Ipke Wachsmuth. Gaze and speech as multimodal input for reference resolution in cooperative tasks. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 84–91. ACM, 2014.

Yusuke Sugano and Andreas Bulling. Seeing with humans: Gaze-assisted neural image captioning. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 148–156, 2016.

Vildan Tanriverdi and Robert JK Jacob. Interacting with eye movements in virtual environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 265–272. ACM, 2000.

Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Object referring in videos with language and human gaze. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4129–4138, 2018.

Petra Wagner, Lars Schillingmann, and Stefan Kopp. Multimodal cues in instructional interactions: The role of gaze, gesture, and prosody. In *Proceedings of the 25th ACM International Conference on Multimodal Interaction*, pages 123–132. ACM, 2023.

# A Mention Linking Details

## A.1 Prompt Schema

The GPT model is constrained to output a JSON object with the following fields:

- `start`, `len`: token index and length of the mention span
- `ingredient`: string or `null`
- `match_type` $\in$ {`exact`, `synonym`, `hypernym`, `brand`, `coref`, `desc`, `object`, `none`}
- `confidence` $\in [0, 1]$
- `antecedent_start`, `antecedent_len`, `antecedent_text` (optional, for coreference)

### A.2 Coreference Rules

Pronouns and demonstratives (*it, this, that, them, these, those*) are linked to the nearest plausible prior mention within approximately 25 tokens, guided by cooking actions. Examples include:

- *open* $\rightarrow$ jar or can
- *drain* $\rightarrow$ pasta or pot
- *put on the stove* $\rightarrow$ pan or pot
- *cut* $\rightarrow$ food item or package

Ambiguous cases default to `ingredient = null, match_type = none`.

### A.3 Post-processing Steps

1. **Normalization:** canonicalize predicted strings to recipe or distractor names; non-matches become `None`.
2. **Enrichment:** add common object aliases (e.g., *hob/cooktop* $\rightarrow$ *stove*) and mark `is_object`.
3. **Timing and surface:** attach token indices and timestamps (`start_ns`, `end_ns`, `start_s`, `end_s`) and surface text.
4. **Mention graph:** assign `mention_id` and `chain_id`, infer `mention_type` (including a small utensil lexicon), and link `antecedent_id` for coreference.

### A.4 Automated Checks

We flag the following cases:

- unmapped ingredients or objects,
- missing antecedents for coreference mentions,
- non-monotone or overlapping timestamps.

These rules ensure consistent and temporally grounded mentions for each recording `ParXrecY`.