Multimodal Speech Recognition for Language-Guided Embodied Agents

Allen Chang[1,2], Xiaoyuan Zhu[1,2], Aarav Monga[1,2], Seoho Ahn[1,2]
Tejas Srinivasan[1], Jesse Thomason[1]

[1]Department of Computer Science, USC, Los Angeles
[2]Center for Artificial Intelligence in Society, USC, Los Angeles

INTERSPEECH 2023
20th-24th August 2023 - Dublin, Ireland

USC

Agents for Instruction Following

Navigation

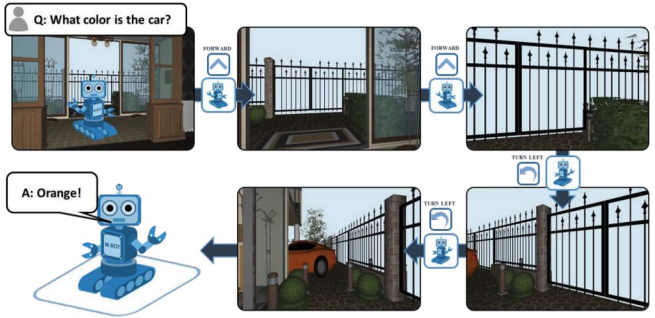
Instruction: Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.

Room-to-Room (Anderson et al., 2018).

Manipulation

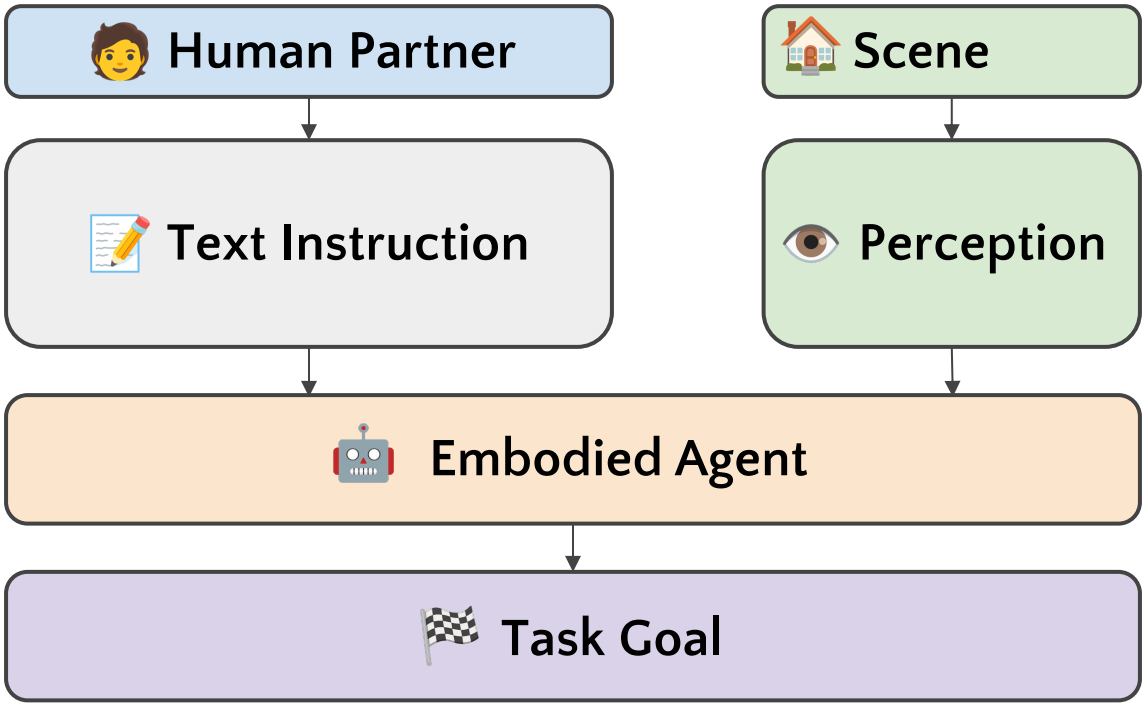
ALFRED (Shridhar et al., 2020).

Question Answering

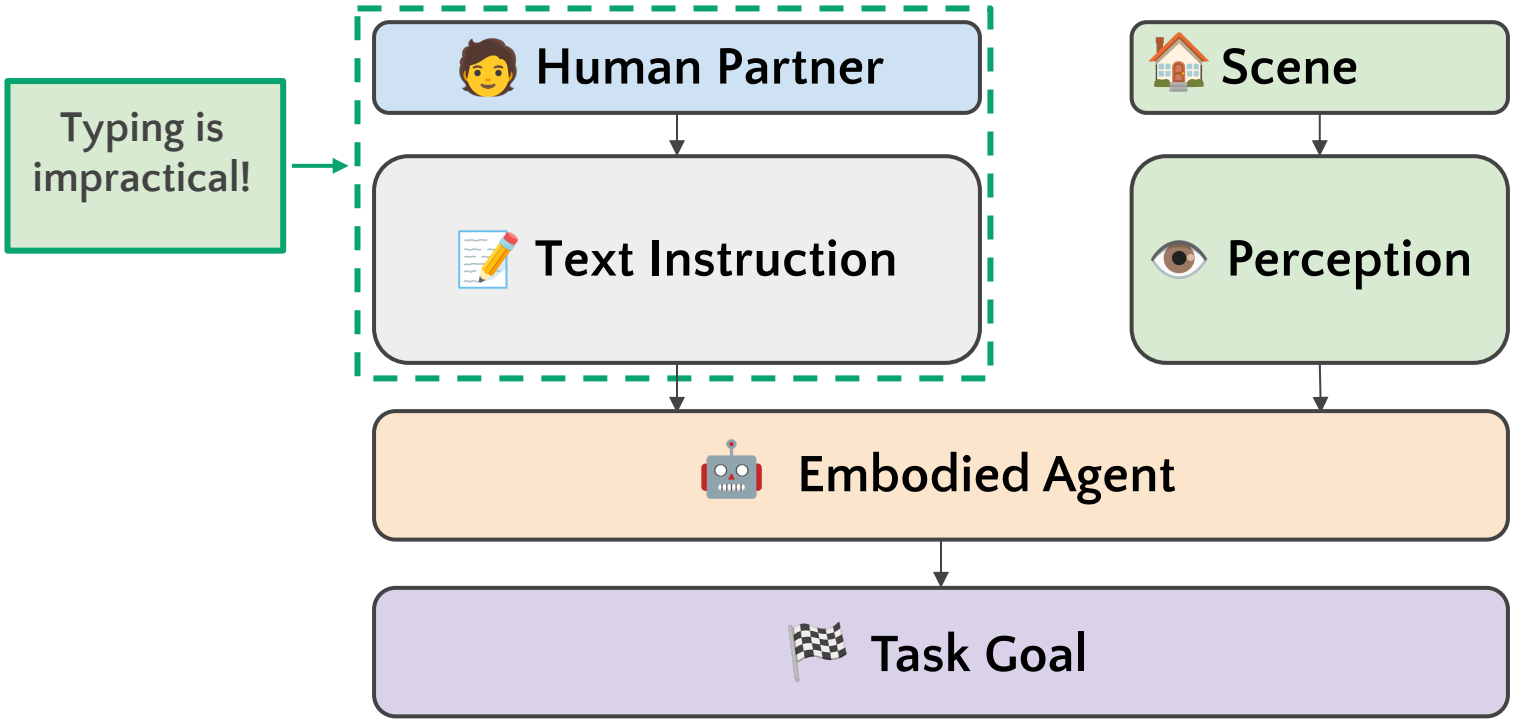


Embodied Question Answering (Das et al., 2018).

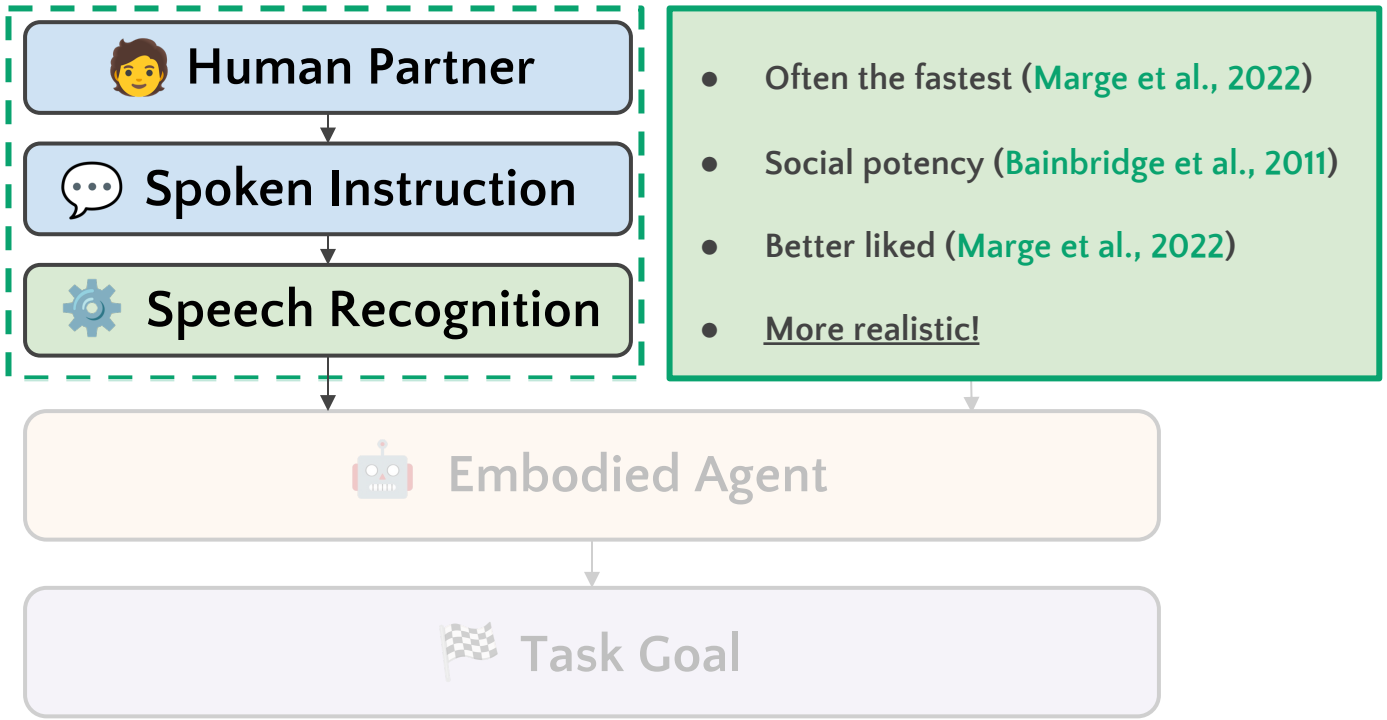
Language-guided embodied agents assume text instructions…



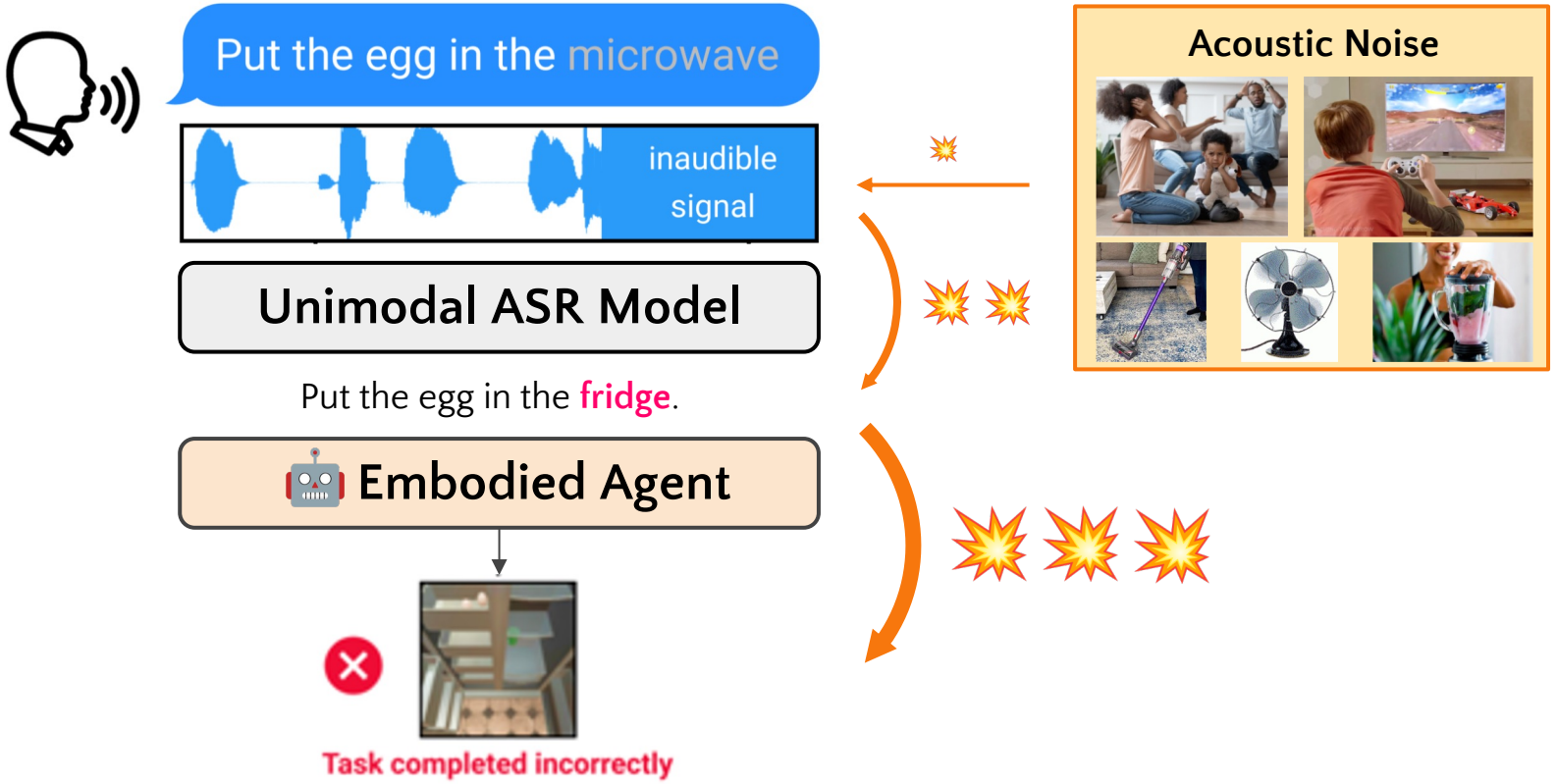
…but in the real world, instructions will be spoken.



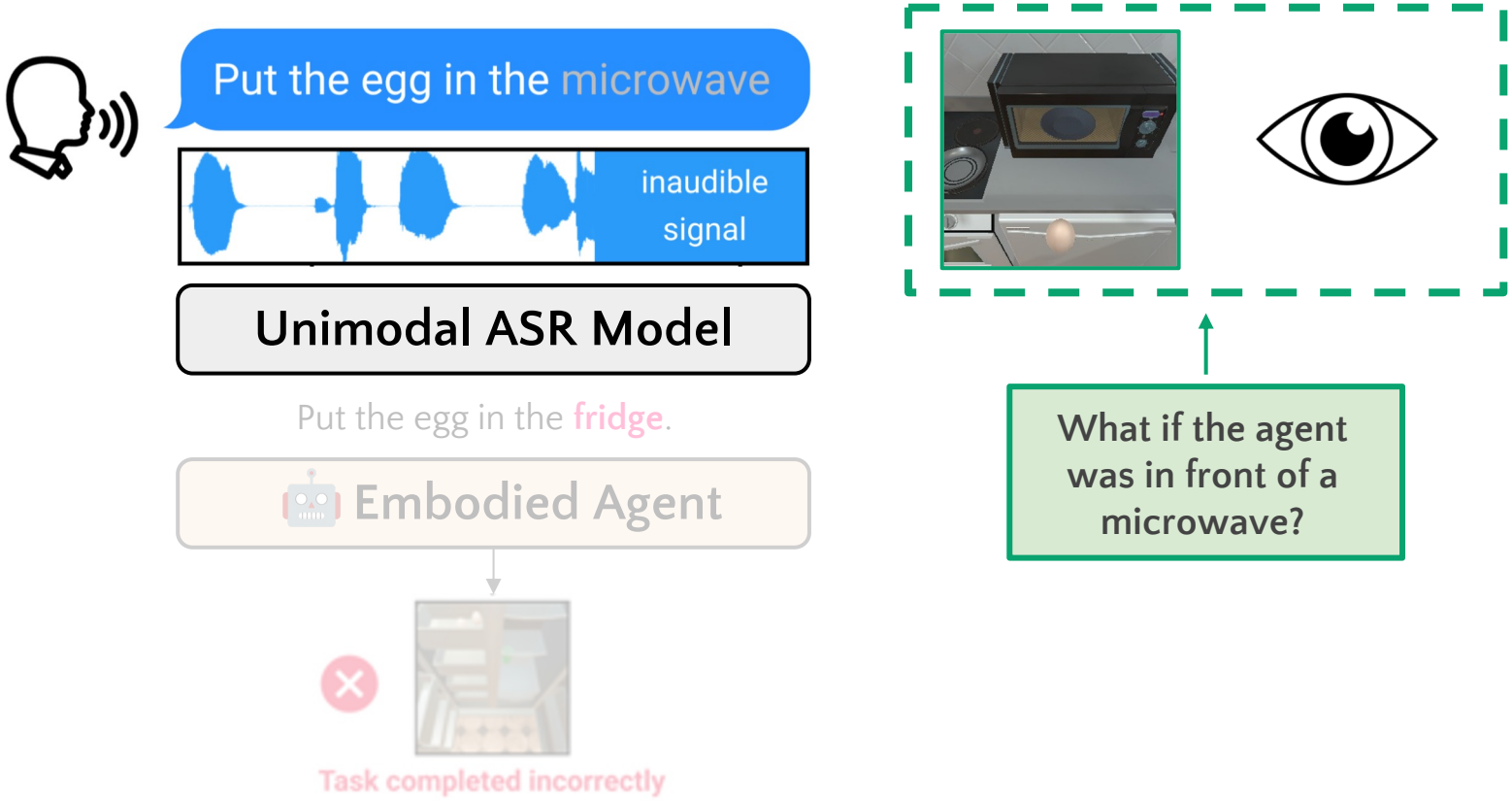
Speech is Better!



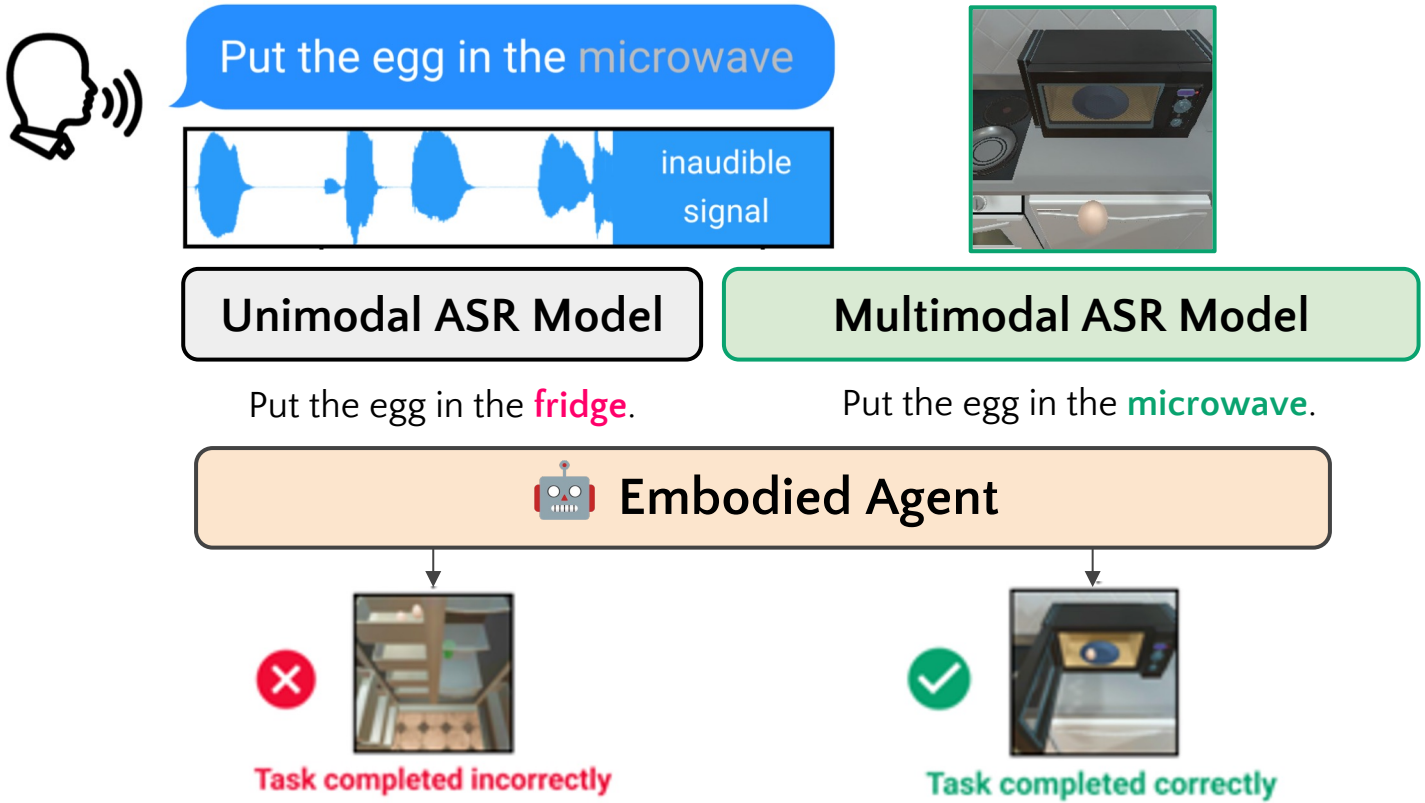
An Obvious Approach. (And Its Failure Mode)



Can Perception Reduce ASR Errors?



Approach: Multimodal ASR for Embodied Agents.



ALFRED Dataset

Goal instruction text G :

"Look at an alarm clock by the light of a lamp."

Sub-goal Text Instructions $I_{1 \ldots K}$:

1 "Go to the table with the safe."

2 "Pick up the clock."

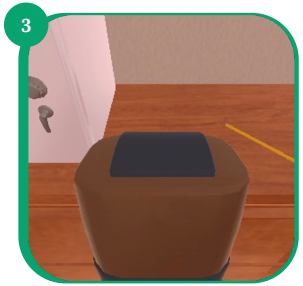
3 "Take the clock to the desk"

... $I_{4 \ldots K}$

Visual observations from human demonstrations of sub-goal tasks $v_{1 \ldots K}$:

1 

2 

3 

... $v_{4 \ldots K}$

Speech-augmented ALFRED Dataset

Goal instruction text G :

"Look at an alarm clock by the light of a lamp."

Sub-goal Text Instructions $I_{1 \ldots K}$:

1 "Go to the table with the safe."

2 "Pick up the clock."

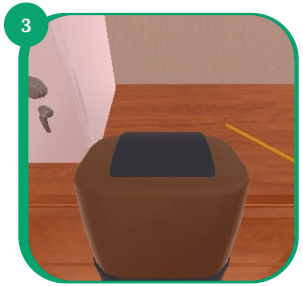
3 "Take the clock to the desk"

... $I_{4 \ldots K}$

Visual observations from human demonstrations of sub-goal tasks $v_{1 \ldots K}$:

1 

2 

3 

... $v_{4 \ldots K}$

1

Extract sub-goal instructions and visual observations

Speech-augmented ALFRED Dataset

Goal instruction text G : "Look at an alarm clock by the light of a lamp."

Sub-goal Text Instructions $I_{1 \ldots K}$:

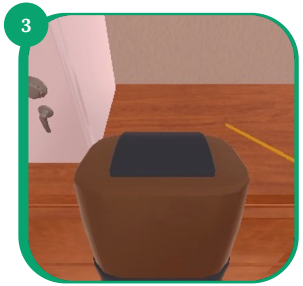
1 "Go to the table with the safe."

2 "Pick up the clock."

3 "Take the clock to the desk"

... $I_{4 \ldots K}$

Visual observations from human demonstrations of sub-goal tasks $v_{1 \ldots K}$:



... $v_{4 \ldots K}$

1 Extract sub-goal instructions and visual observations

2 Apply TTS Model to sub-goal instructions

Speech-augmented ALFRED Dataset

Goal instruction text G : "Look at an alarm clock by the light of a lamp."

Sub-goal Text Instructions $I_{1 \ldots K}$:

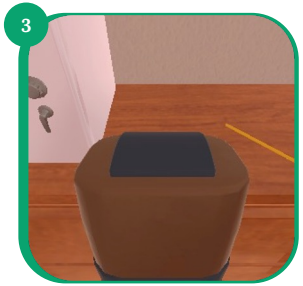
1 "Go to the table with the safe."

2 "Pick up the clock."

3 "Take the clock to the desk"

... $I_{4 \ldots K}$

Visual observations from human demonstrations of sub-goal tasks $v_{1 \ldots K}$:



... $v_{4 \ldots K}$

1 Extract sub-goal instructions and visual observations

2 Apply TTS Model to sub-goal instructions

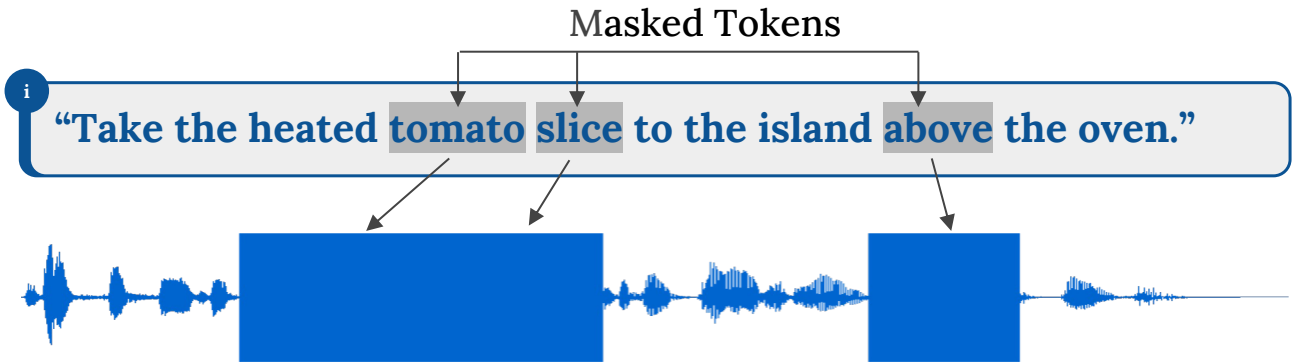
3 Perturb with word-level masking

Audio Masking Policies

Mask Type: Gaussian noise

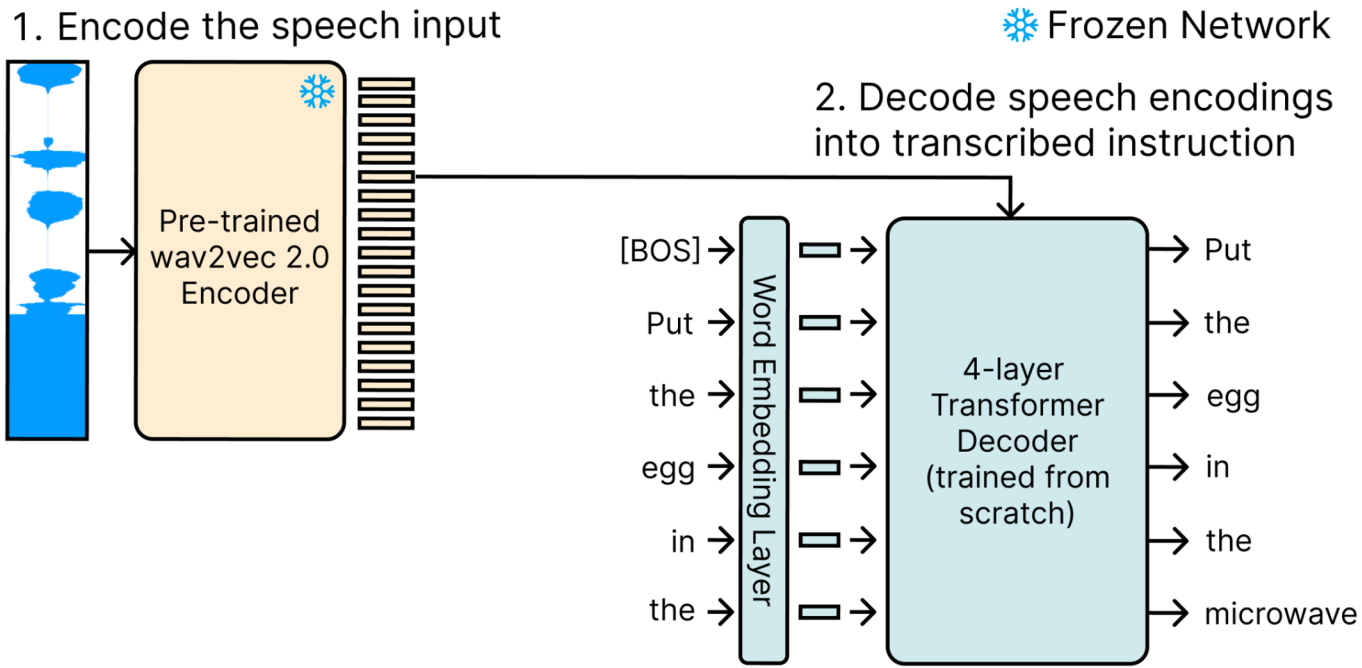
No masking policy: Clean audio
Noun masking policies: 20%, 40%, 100% of all nouns (identified with NLTK)
All word masking policies: 20%, 40% of all words

Ground Truth (N=13):

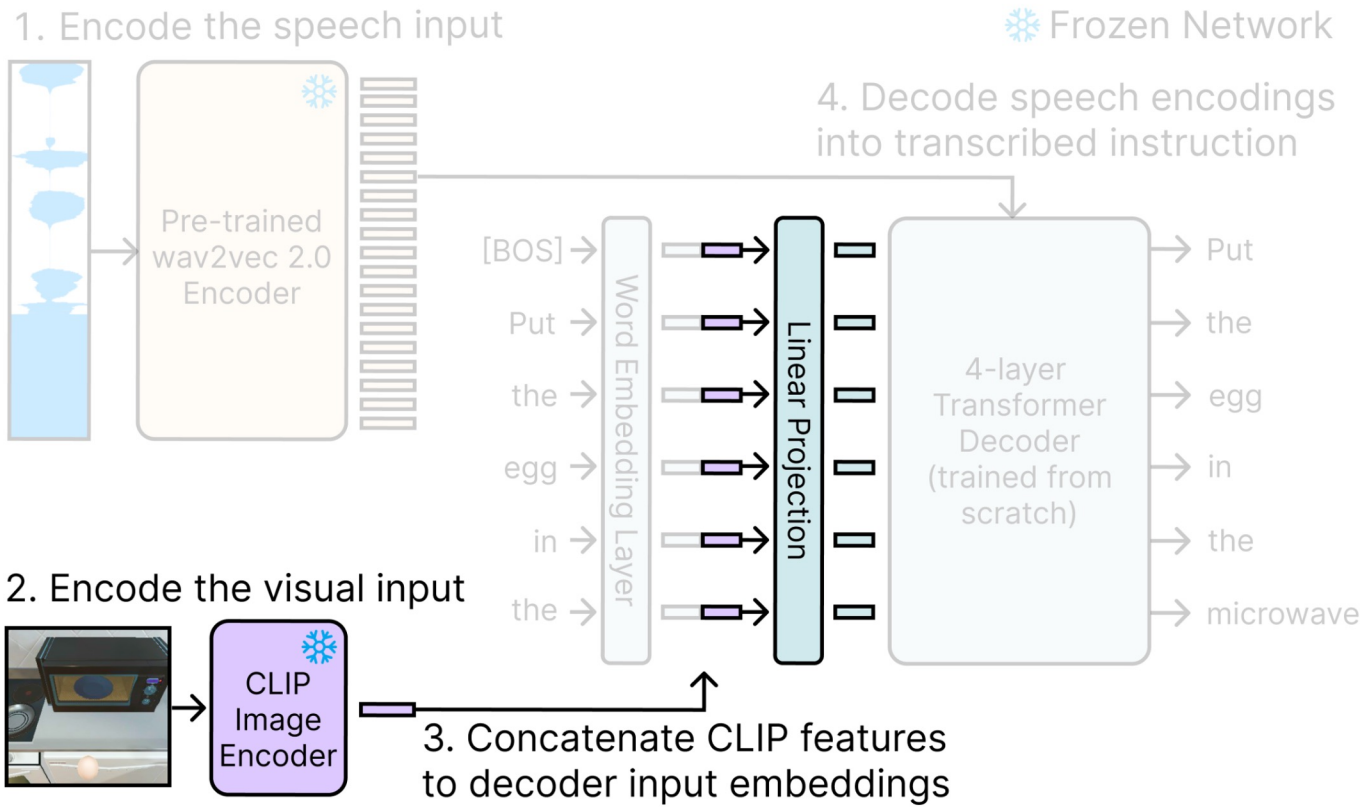


Speech Waveform:

Near-identical Unimodal v.s. Multimodal Architectures



Near-identical Unimodal v.s. Multimodal Architectures

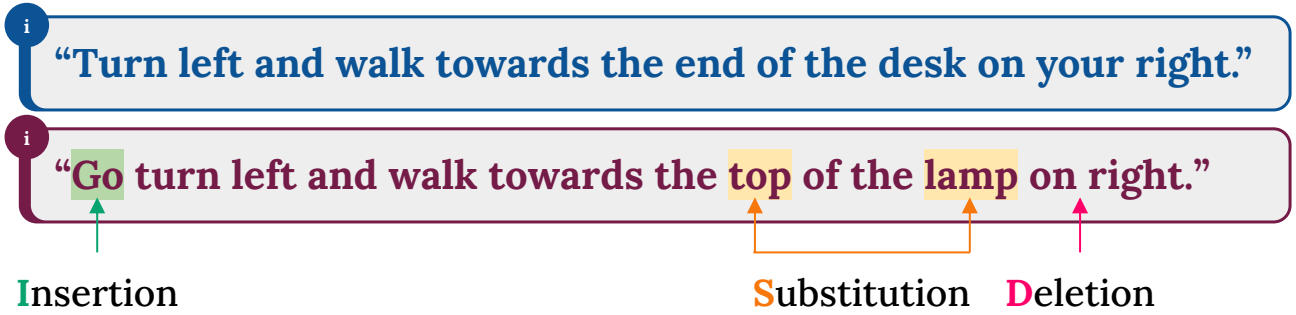


Metrics: ↓ Word Error Rate (WER)

Word Error Rate: % degree of word-level error

Ground Truth (N=13):

ASR Prediction:



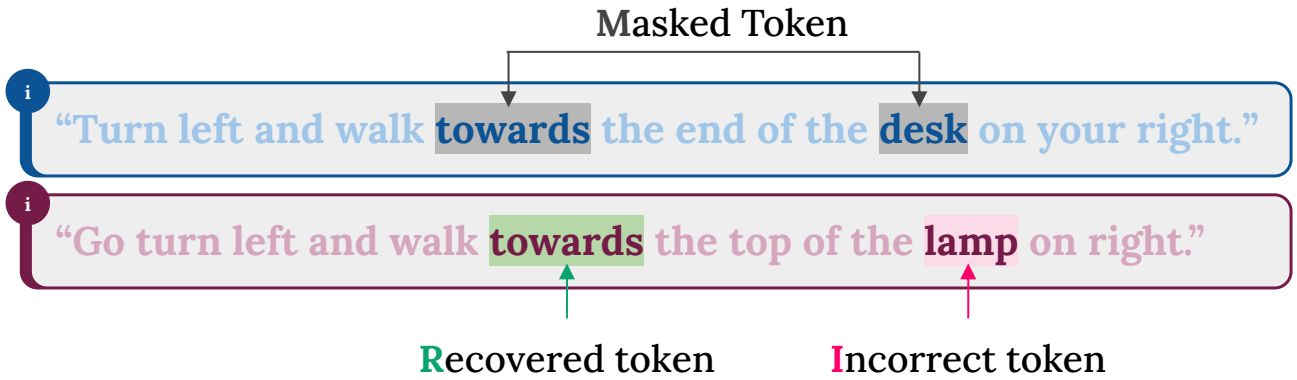
$$\text{WER} = \frac{\#\ \text{Insertions} + \#\ \text{Substitutions} + \#\ \text{Deletions}}{\#\ \text{of tokens in Ground Truth (N)}}$$

Metrics: ↑ Recovery Rate (RR)

Recovery Rate: % of masked words recovered

Ground Truth (N=13):

ASR Prediction:



$$\text{RR} = \frac{\text{\# Recovered tokens}}{\text{\# Masked tokens}}$$

Relative Metrics: $\downarrow \Delta_{WER}$ and $\uparrow \Delta_{RR}$

$$\downarrow \Delta_{WER} \text{ Relative Change in WER} = \frac{\text{WER}_{\textit{Multimodal}} - \text{WER}_{\textit{Unimodal}}}{\text{WER}_{\textit{Unimodal}}} \times 100\%$$

$$\uparrow \Delta_{RR} \text{ Relative Change in RR} = \frac{\text{RR}_{\textit{Multimodal}} - \text{RR}_{\textit{Unimodal}}}{\text{RR}_{\textit{Unimodal}}} \times 100\%$$

Results

Is multimodal ASR helpful for spoken instruction recognition?

Multimodal ASR Reduces Error and Improves Recovery

Multimodal > Unimodal for all audio masking policies.

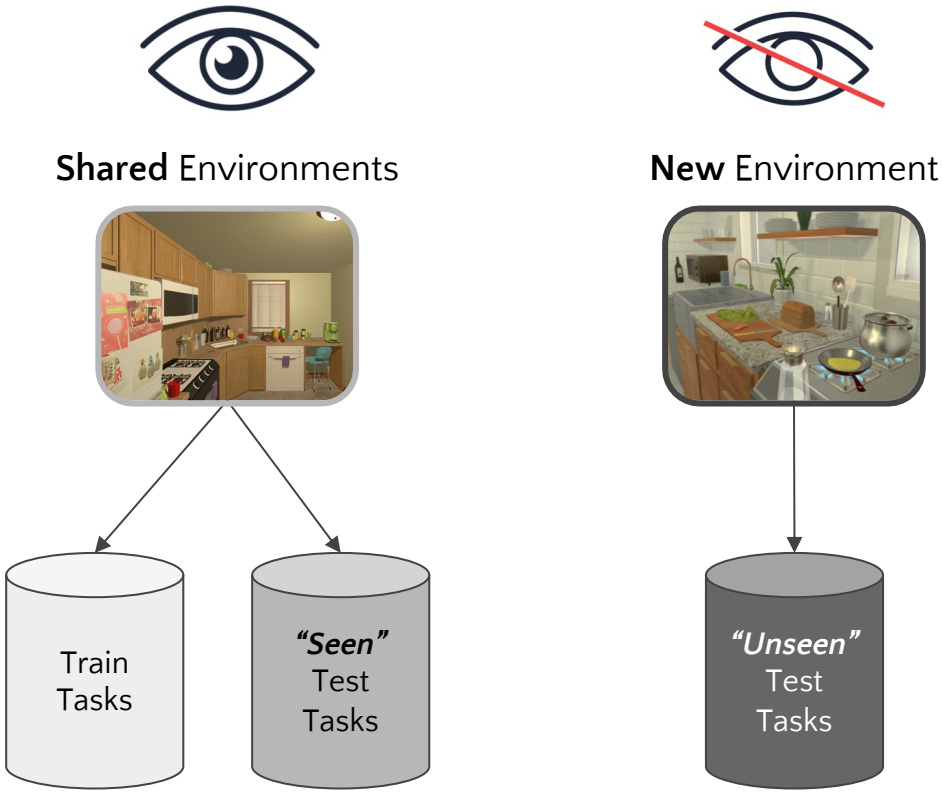
ASR Model	No Mask	Only Nouns			All Words	
		20%	40%	100%	20%	40%
Word Error Rate ↓						
Unimodal	12.6	20.0	26.4	34.0	35.2	49.9
Multimodal	**11.9**	**19.9**	**24.5**	**30.4**	**29.3**	**46.4**
Recovery Rate ↑						
Unimodal	–	61.1	56.1	48.0	51.4	38.8
Multimodal	–	**64.3**	**60.5**	**56.8**	**57.7**	**45.9**

Results

Does multimodal ASR generalize to …

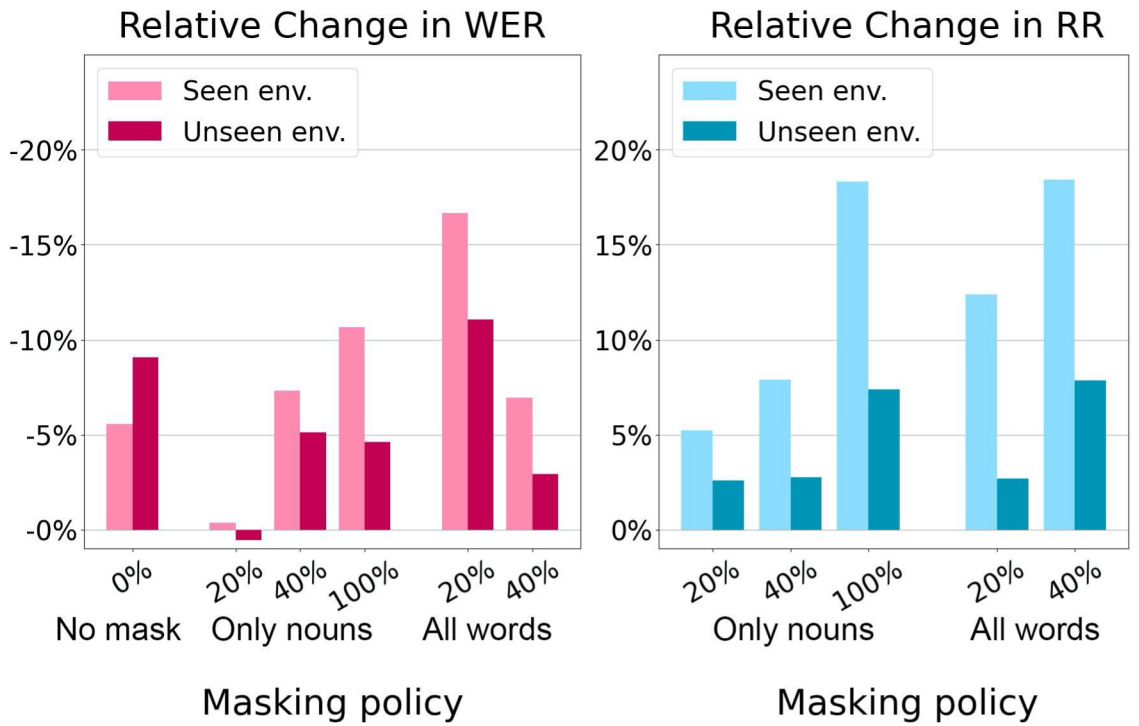
… unfamiliar visual signals?
… unfamiliar audio signals?

Does multimodal ASR generalize to new unseen environments?



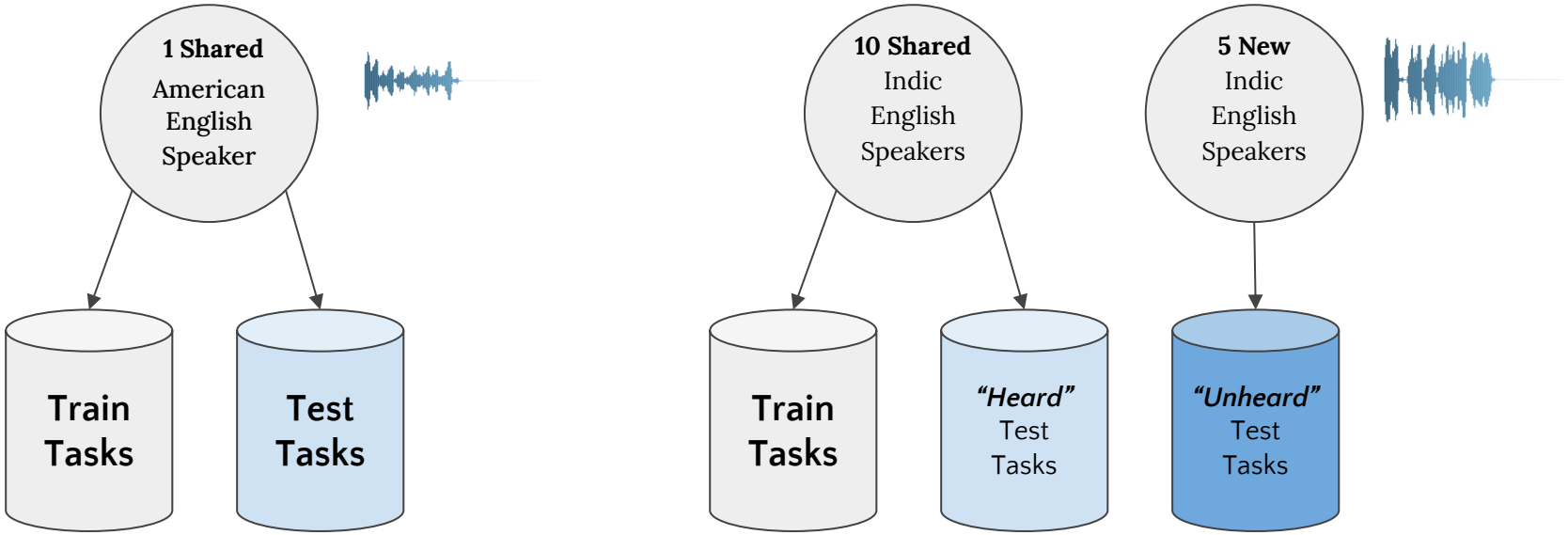
Yes, Multimodal ASR Generalizes to Seen Environments

- **Multimodal > Unimodal for seen and unseen environments.**
- **Visuals are more helpful in “Seen"**



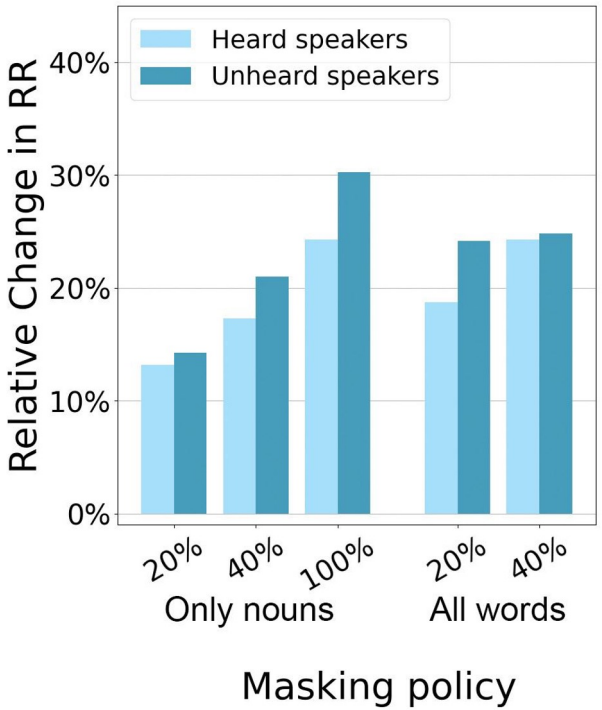
Does multimodal ASR generalize to new unheard speakers?

> "Turn left and walk towards the end of the desk on your right."



Yes, and Multimodal ASR is More Helpful for Unheard Speakers

- Multimodal > Unimodal for heard and unheard speakers.
- Visuals are more helpful in “Unheard"



Results

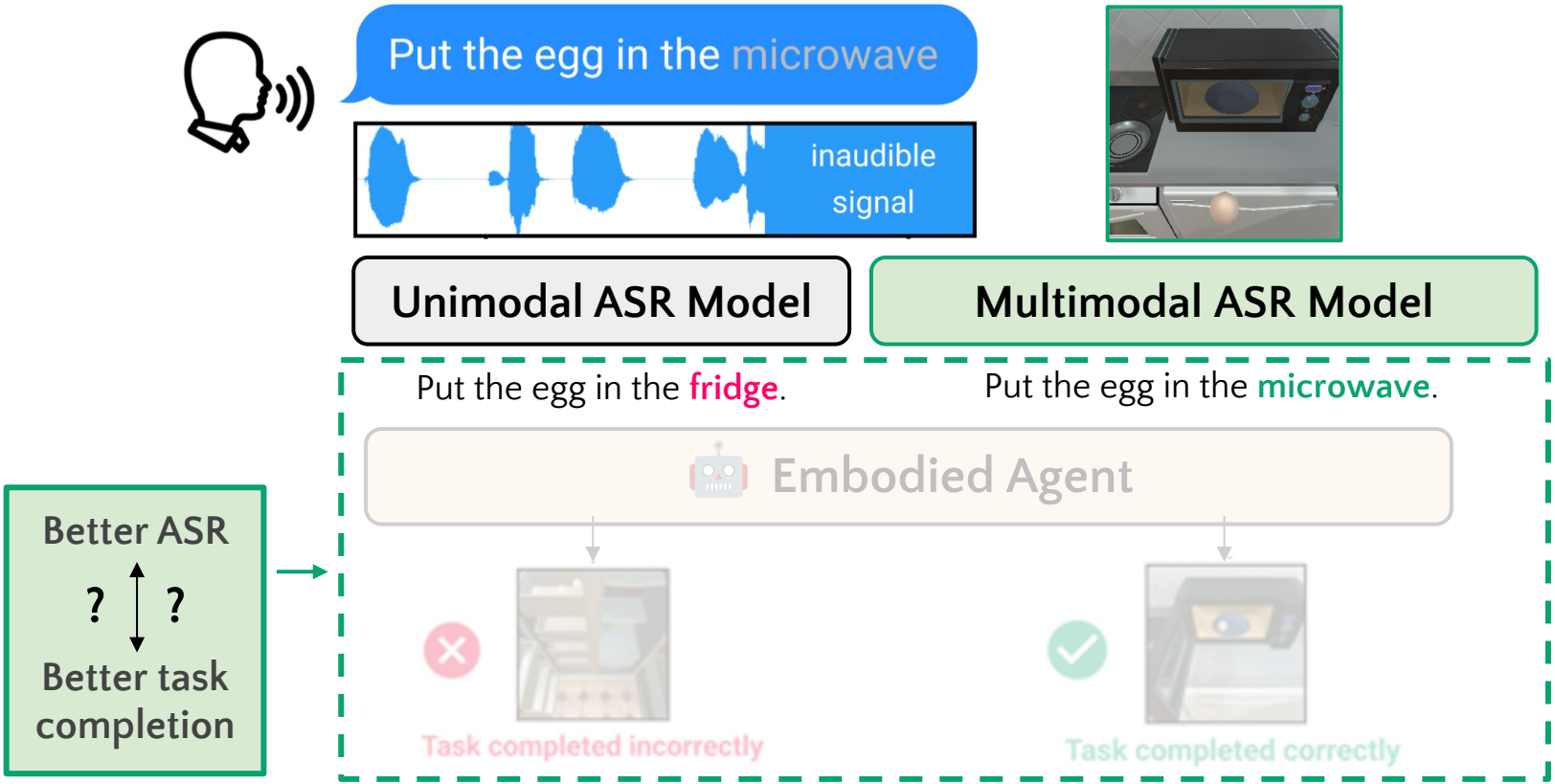
Does multimodal ASR help for the right reasons?

Multimodal ASR Performs Better for Visually Salient Words

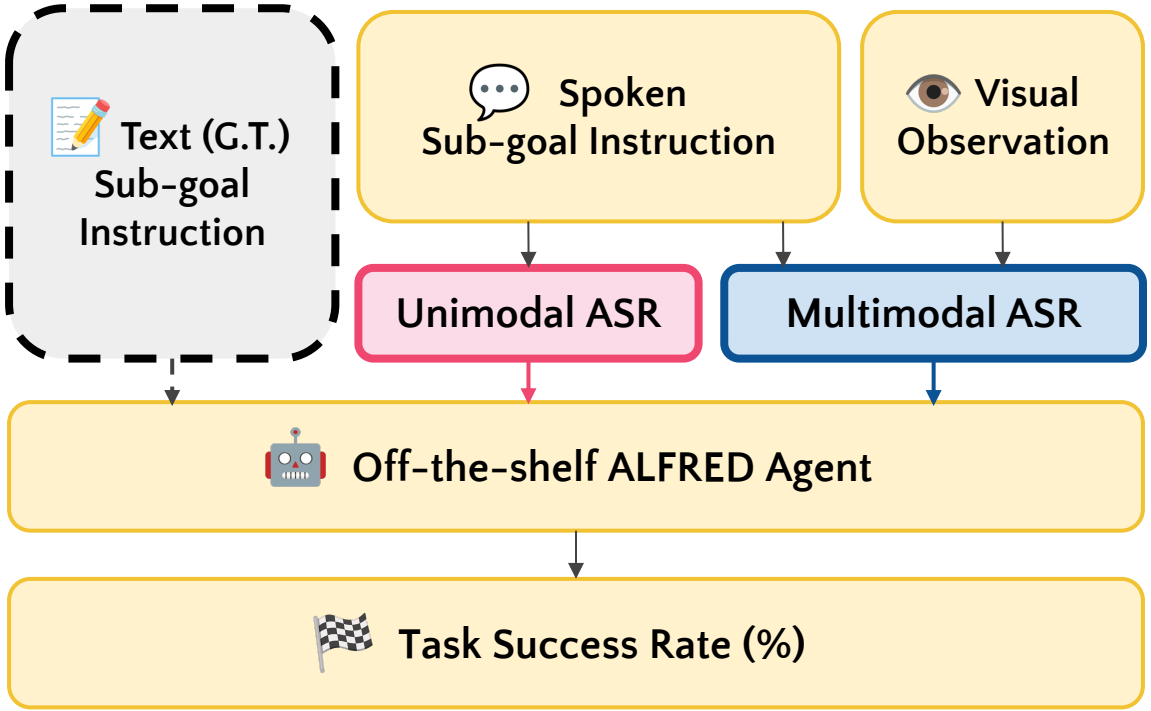
Table 2: Δ_{RR} *on the subset of words corresponding to nouns and non-nouns from the random all-words masking policies.*

Speaker(s)	**POS**	**20% Masking**		**40% Masking**	
		Seen	*Unseen*	*Seen*	*Unseen*
American	Noun	**+12.7%**	**+03.0%**	**+31.0%**	**+14.6%**
American	Other	+04.4%	−02.2%	−00.1%	−01.5%
Indic (Heard)	Noun	**+20.1%**	**+05.5%**	**+40.1%**	**+18.8%**
Indic (Heard)	Other	−04.3%	−11.9%	+02.5%	−01.8%
Indic (Unheard)	Noun	**+25.7%**	**+07.9%**	**+45.5%**	**+24.7%**
Indic (Unheard)	Other	−00.7%	−08.1%	−01.6%	−00.2%

Multimodality is helpful for ASR. Does it help agents complete tasks?

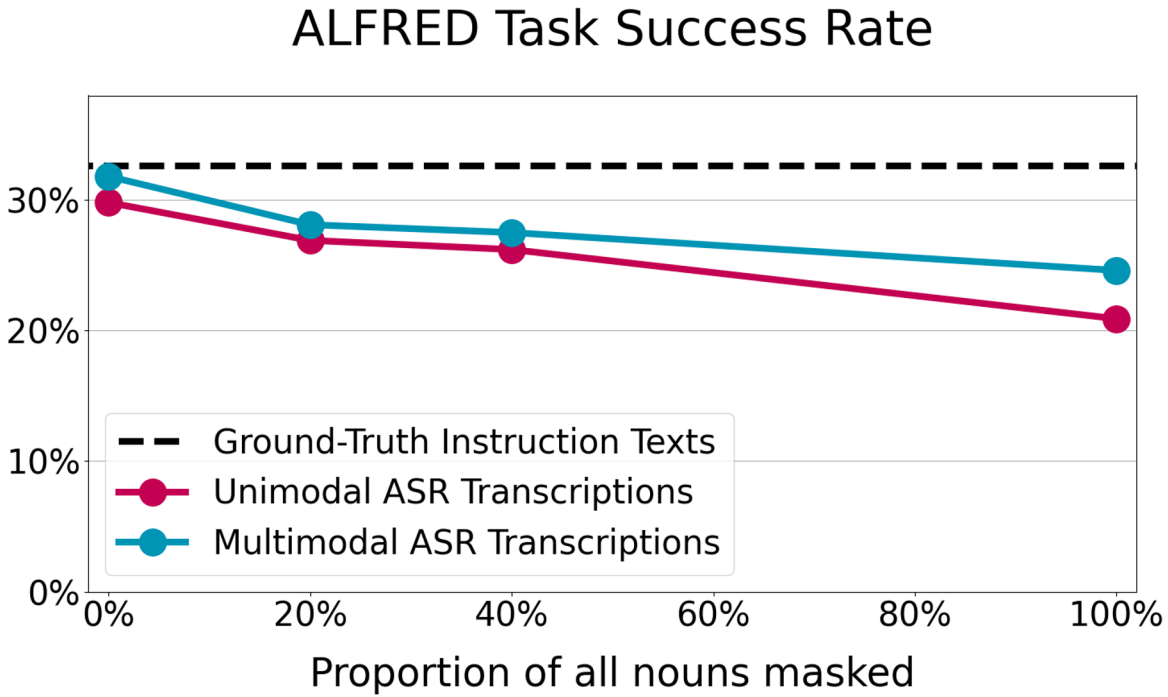


Agents receive G.T., unimodal, and multimodal text.



Multimodality can help agents complete tasks!

- Both ASR methods perform worse than with the text instructions
- Multimodal ASR achieves higher task success rate than unimodal ASR

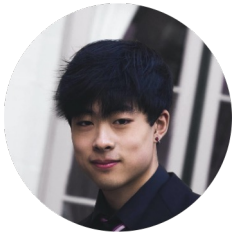


USC

SCAN ME

Multimodal Speech Recognition For Language-Guided Embodied Agents

Paper ID: 2262



Allen Chang

Xiaoyuan Zhu

Aarav Monga

Seoho Ahn

Tejas Srinivasan

Jesse Thomason

Takeaways:

Problem: Agents following spoken instructions are influenced by errors in ASR.

Our work finds that:

1. An embodied agent with ASR can use their visual observations to be more robust.
2. Multimodal ASR can enhance spoken instruction recognition from heard and unheard speakers, even in unseen environments.
3. Reducing ASR errors can also reduce downstream task-completion failure.