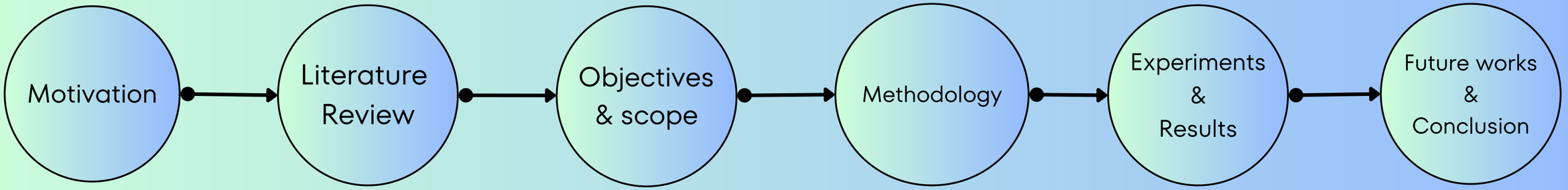
Extract Human Preferences From Language Inputs Using GPT

Supervisor:
Associate Professor Ang Wei Tech

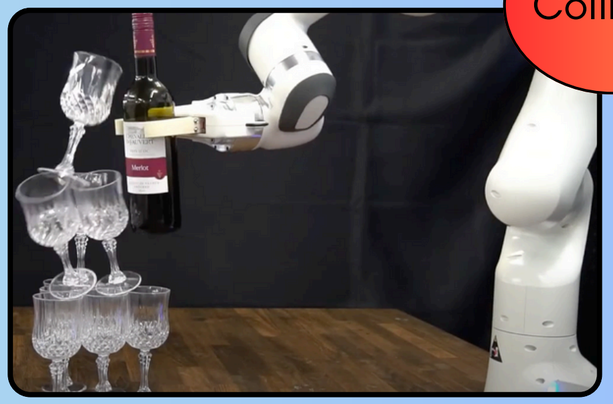
Presenter:
Mohamed Raizee

Agenda


Motivation
Literature Review
Objectives & scope
Methodology
Experiments & Results
Future works & Conclusion

Motivation

- Improve human-robot interaction by using language and vision
- Grounding of language in robotics
- Modify predefined trajectory to avoid collision with obstacles
- Adjustable trajectory based on intensities



[1]

Without trajectory modification

[1]

Reshape initial robotic arm trajectory

Trajectory

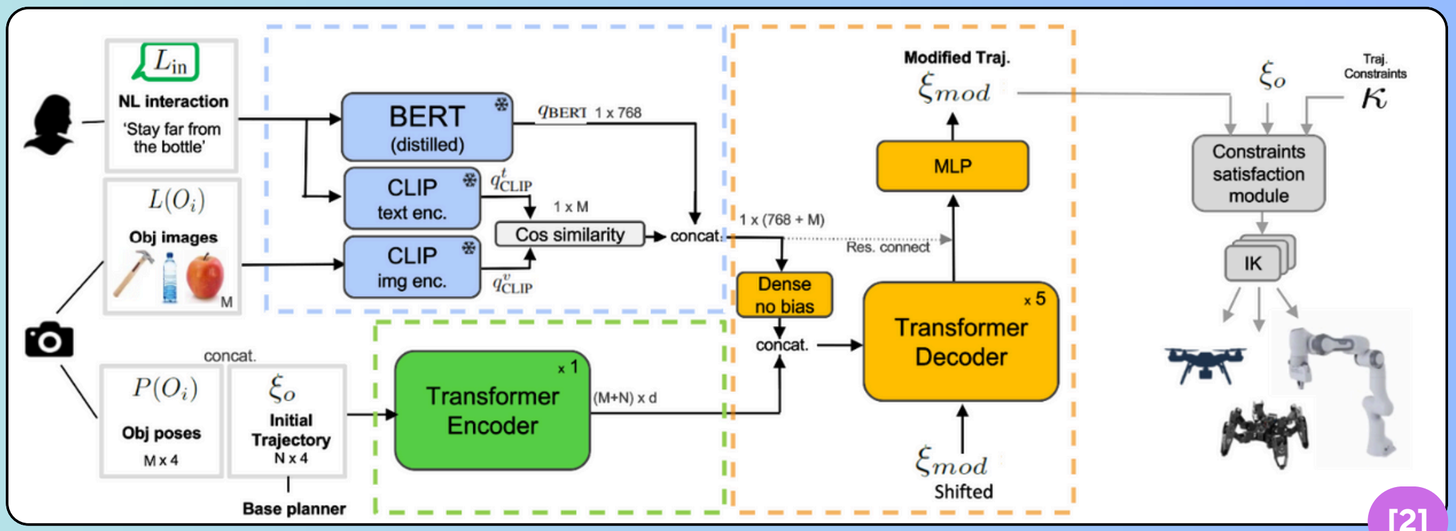
[1]

- Initial Trajectory
- Low intensity trajectory modification
- High intensity trajectory modification

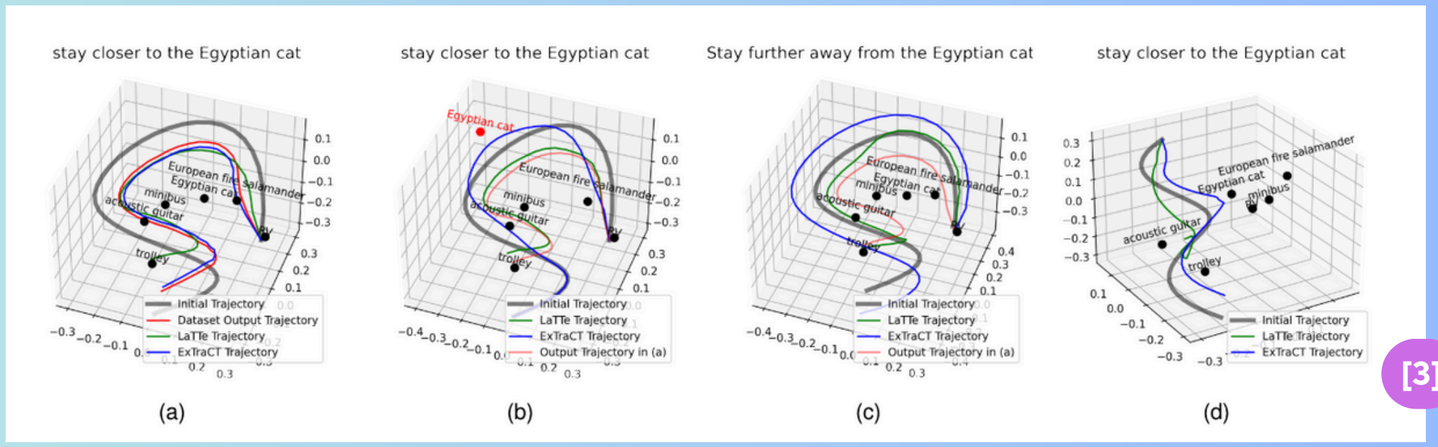
Literature Review

- LAnguage Trajectory TransformEr
- End-to-end training method
- Unable to generalise as well due to the nature of the training
- Requires large amount of training data

LATTE Architecture



[2]

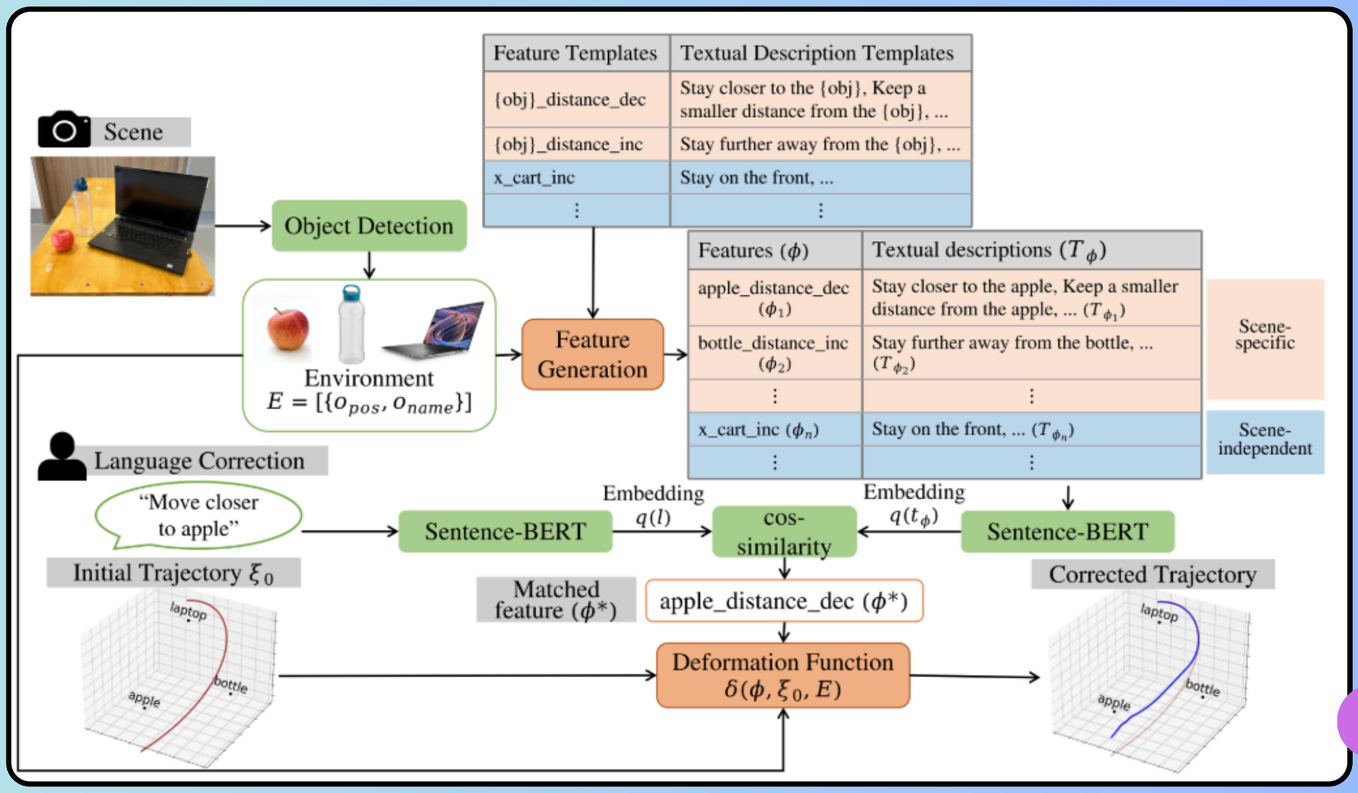


[3]

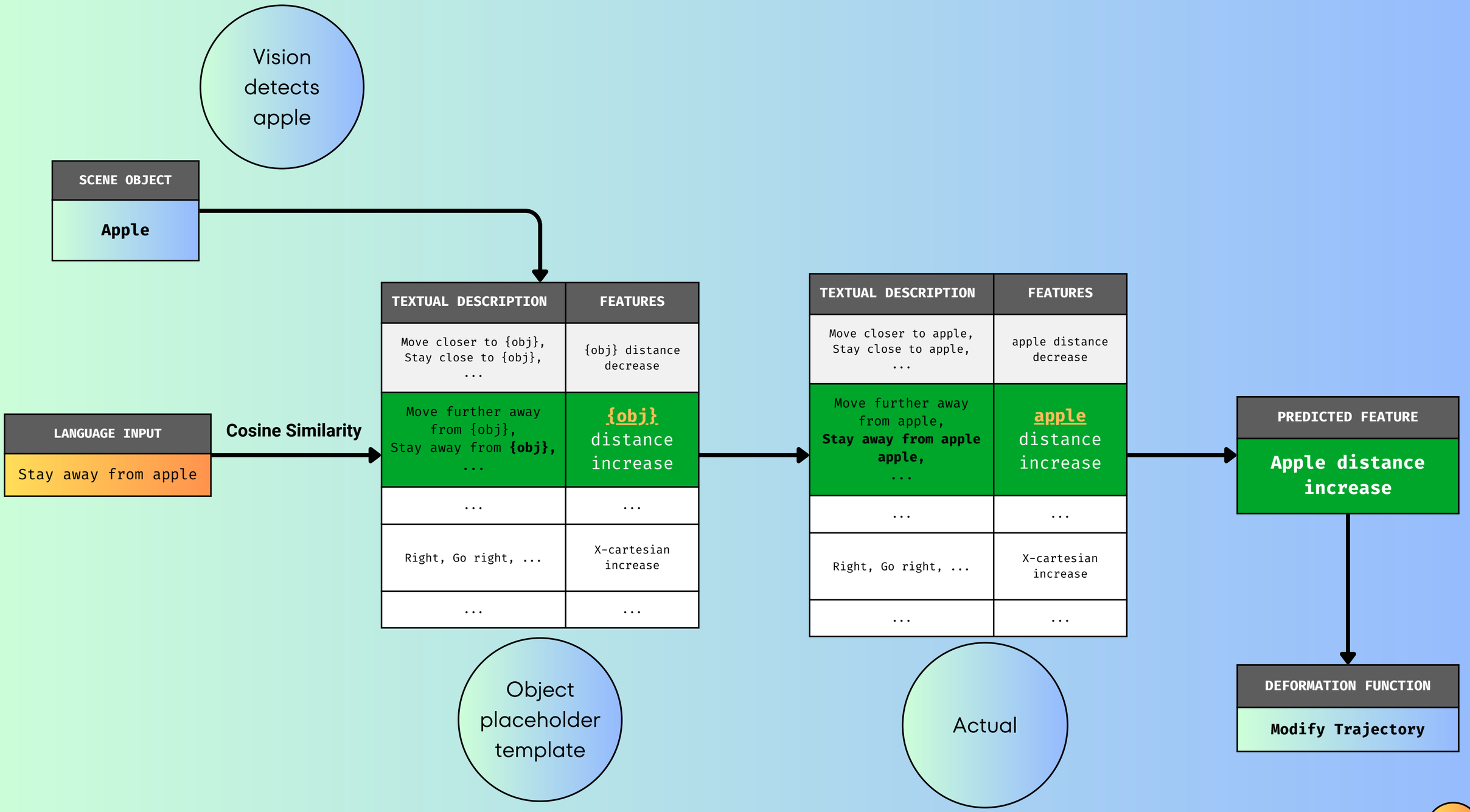
ExTraCT Architecture

Segments

- **Vision:** Object detection model to detect object in scenes
- **Language:** BERT sentence similarity model to handle language command
- **Feature Generation:** Textual description to features mapping to parse into deformation function for trajectory modification



[3]


Vision detects apple
SCENE OBJECT
Apple
LANGUAGE INPUT
Stay away from apple
Cosine Similarity
TEXTUAL DESCRIPTION
FEATURES
Move closer to {obj}, Stay close to {obj}, ...
{obj} distance decrease
Move further away from {obj}, Stay away from {obj}, ...
{obj} distance increase
...
...
Right, Go right, ...
X-cartesian increase
...
...
Object placeholder template
TEXTUAL DESCRIPTION
FEATURES
Move closer to apple, Stay close to apple, ...
apple distance decrease
Move further away from apple, Stay away from apple apple, ...
apple distance increase
...
...
Right, Go right, ...
X-cartesian increase
...
...
Actual
PREDICTED FEATURE
Apple distance increase
DEFORMATION FUNCTION
Modify Trajectory

Limitation

- Suboptimal accuracy for single command
- Inability to handle multiple consecutive commands (chained)
- Manual keyboard input for language command

Definition

Command type	Example
Single	Get away from bottle
Chained	Get away from bottle **and** keep left, ...

Objectives

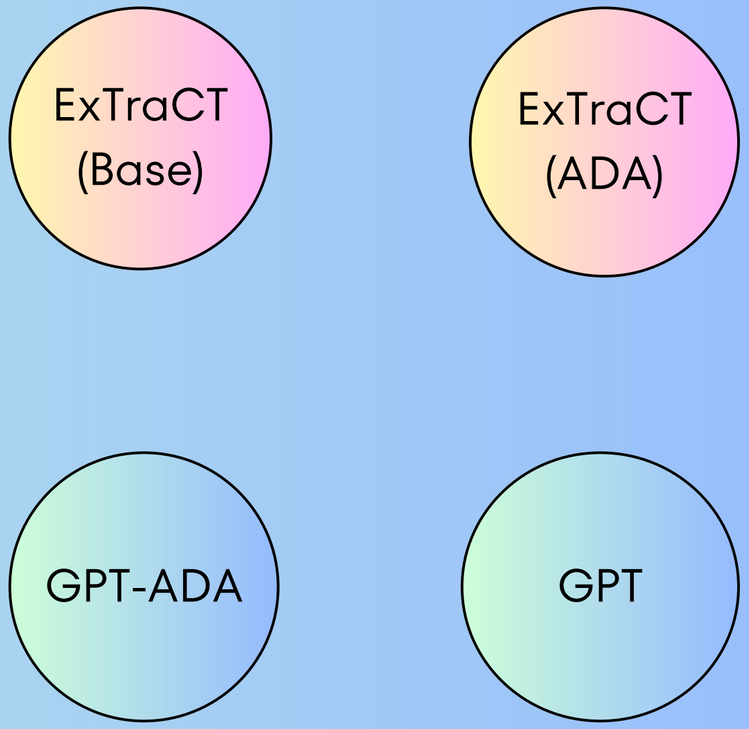
- Improve performance for feature classification
- Increase robustness for consecutive chained commands
- Incorporate speech-to-text capability
- Integrate overall pipeline with robotic arms

Scope

- Enhancements limited to language side, vision excluded
- Generic manipulation features used
  - Object distance related feature
  - Cartesian distance related feature (x,y,z axes)
- Models used are from OpenAI

Methodology

- ExTraCT (Base) architecture using BERT embedding model
- Larger embedding model ADA
- GPT for preprocessing & ADA for feature matching
- GPT for preprocessing & feature matching



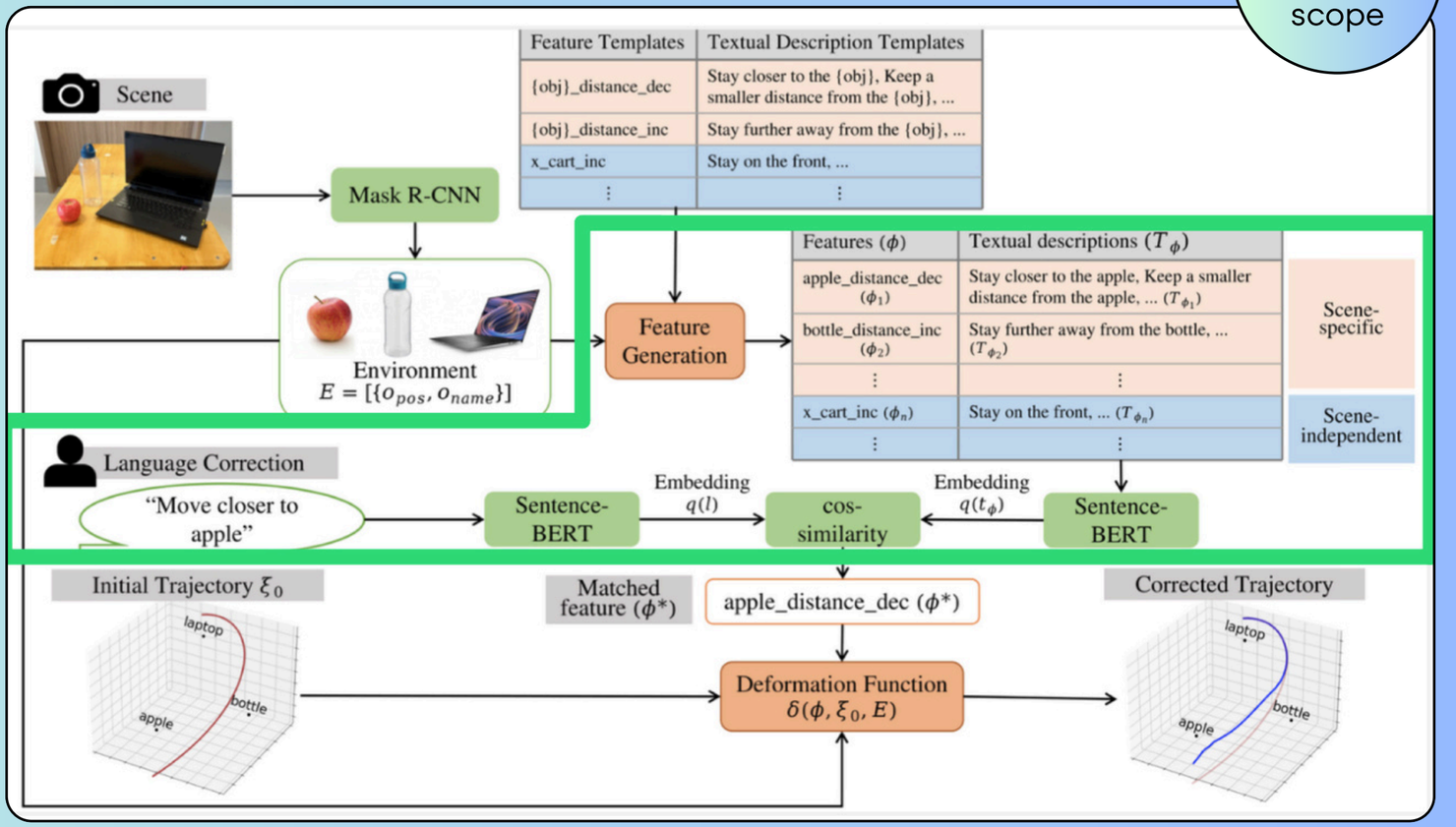
*GPT refers to GPT-4 Turbo chat model [4]

ExTraCT Base

- Base approach from literature review

- Suboptimal accuracy for single command

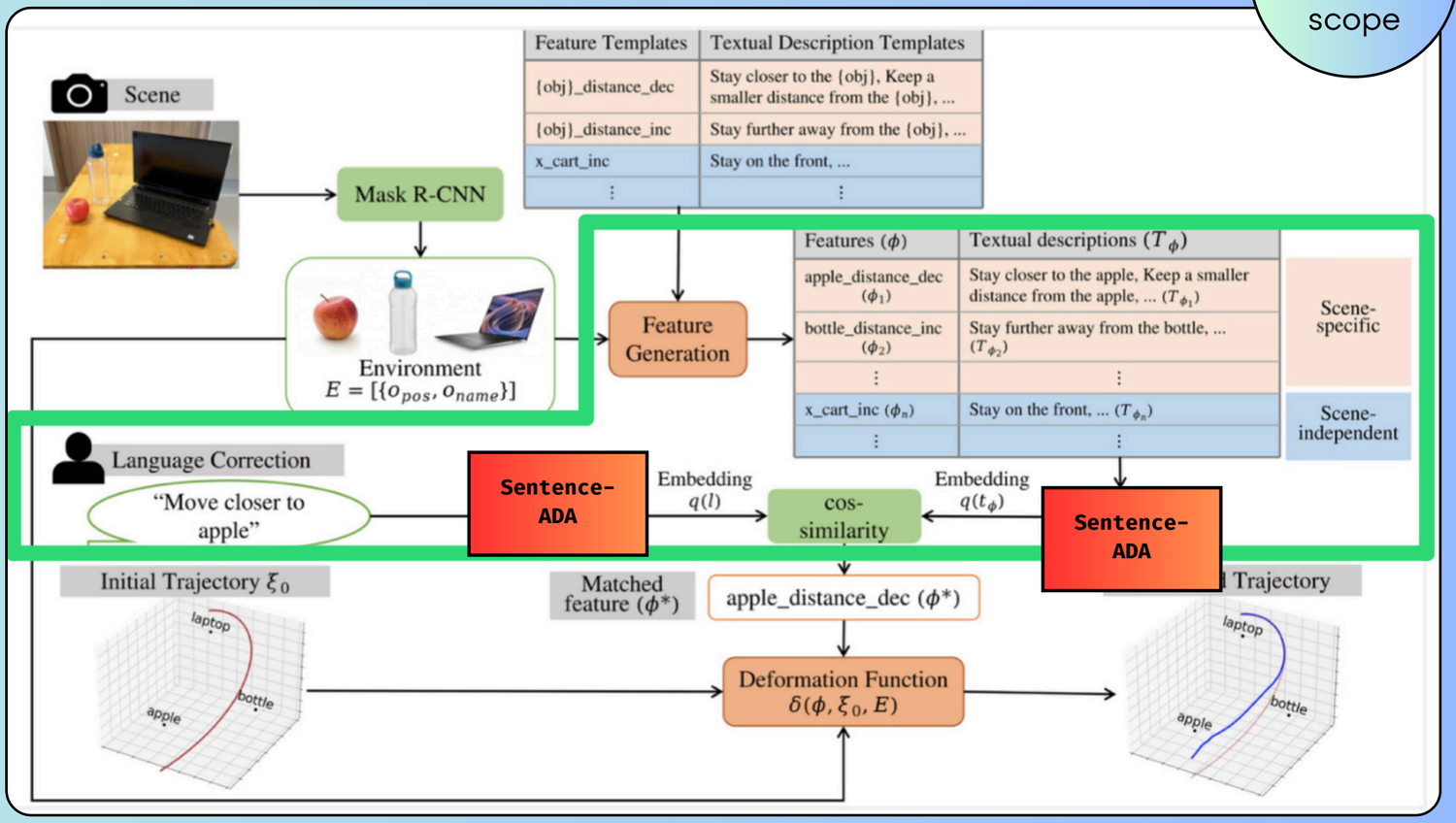
- Inability to comprehend consecutive chained commands



ExTraCT ADA

- Same architecture as baseline
- Larger embedding model

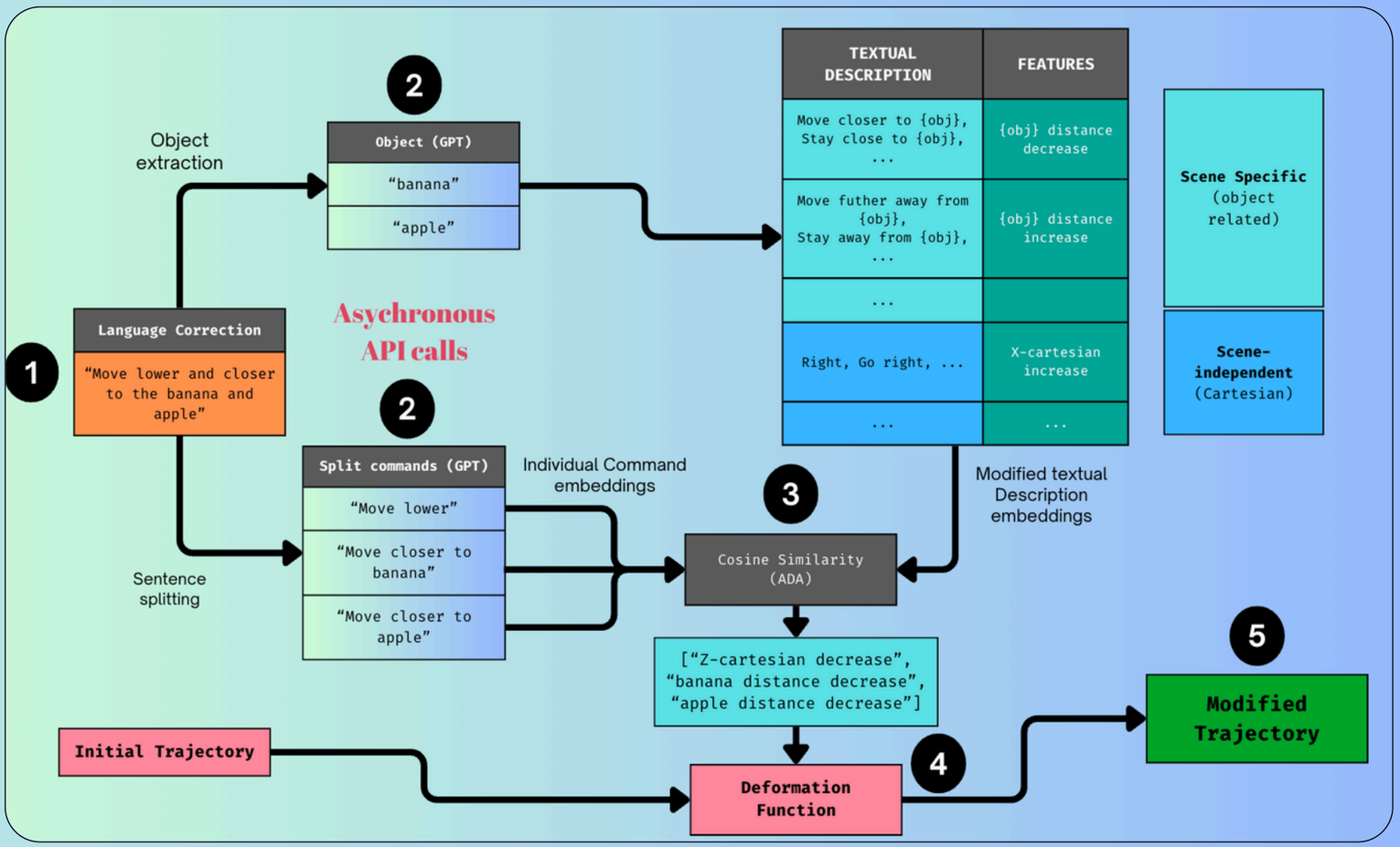
- Perfect accuracy for single command
- Inability to comprehend consecutive chained commands
- Limitations of architecture



GPT-ADA

- Preprocessing steps in (2)
- Divide & conquer
- Feature matching using Ada

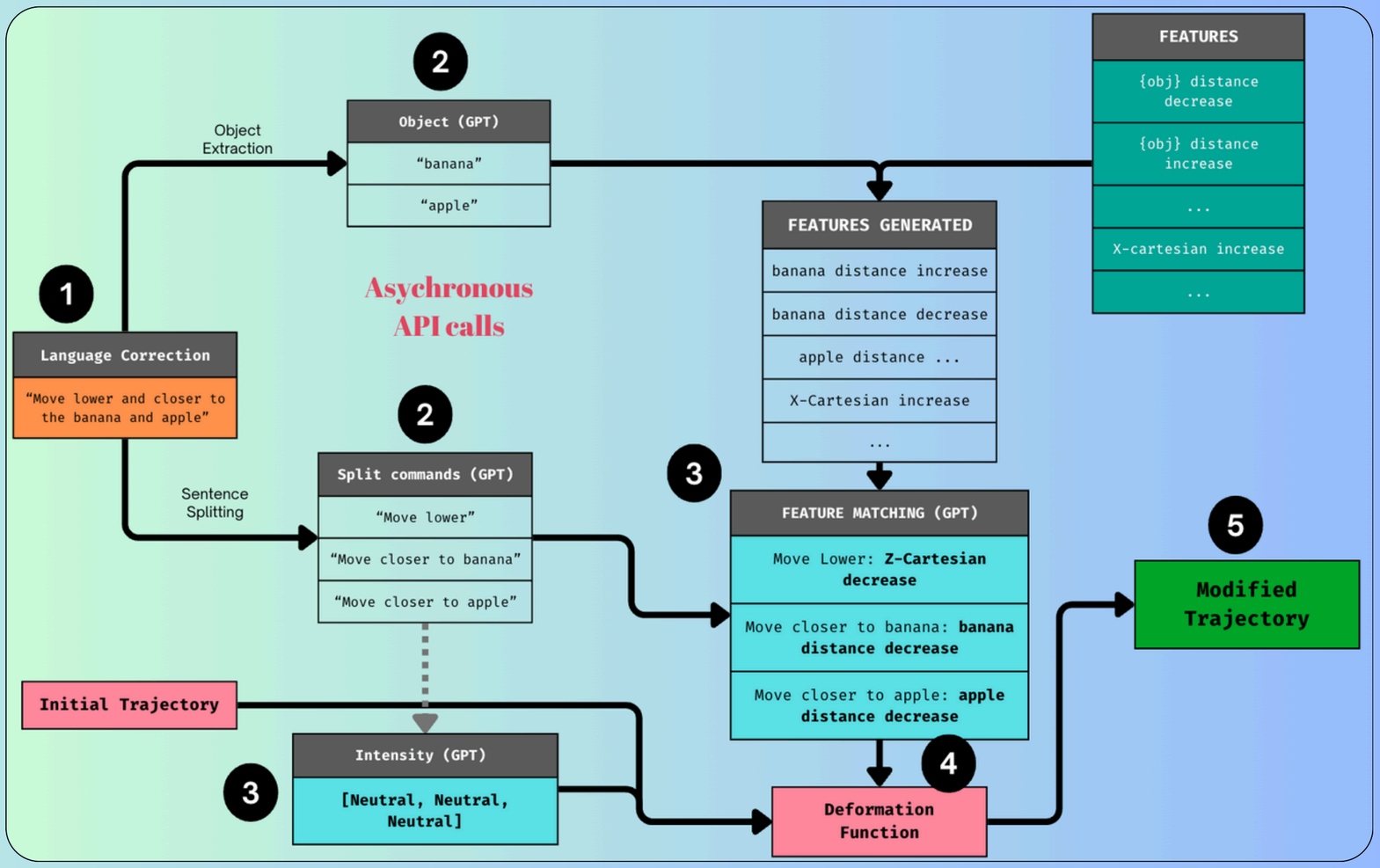
- Perfect accuracy for single command
- Decent chained command accuracy but room for improvement (~ 88%)
- Large use of vector embedding space locally



GPT

- Feature matching using GPT

- Perfect accuracy for single command
- Exceptionally high accuracy for chained command (~ 95%)
- No vector embedding space required



Experiment 1

- Performance accuracy for single command
- Performance accuracy for chained commands
- Select best performing approach to replace overall pipeline

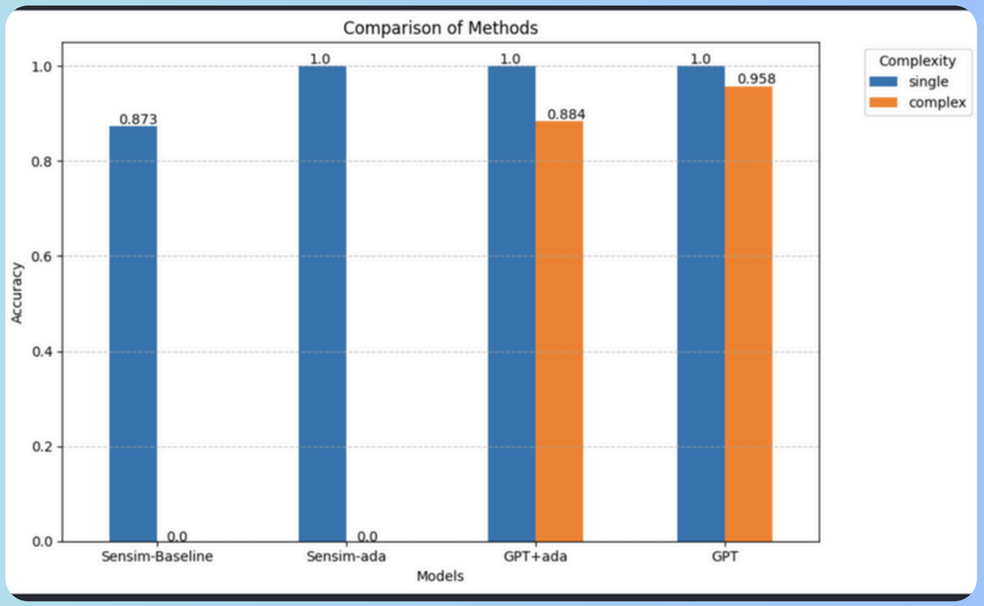
Breakdown	Count
Single command	55
Chained commands	95
Total	150

Single	Complex
Move near the bottle	**Approach the vase and move left then get nearer to the orange and stay down**

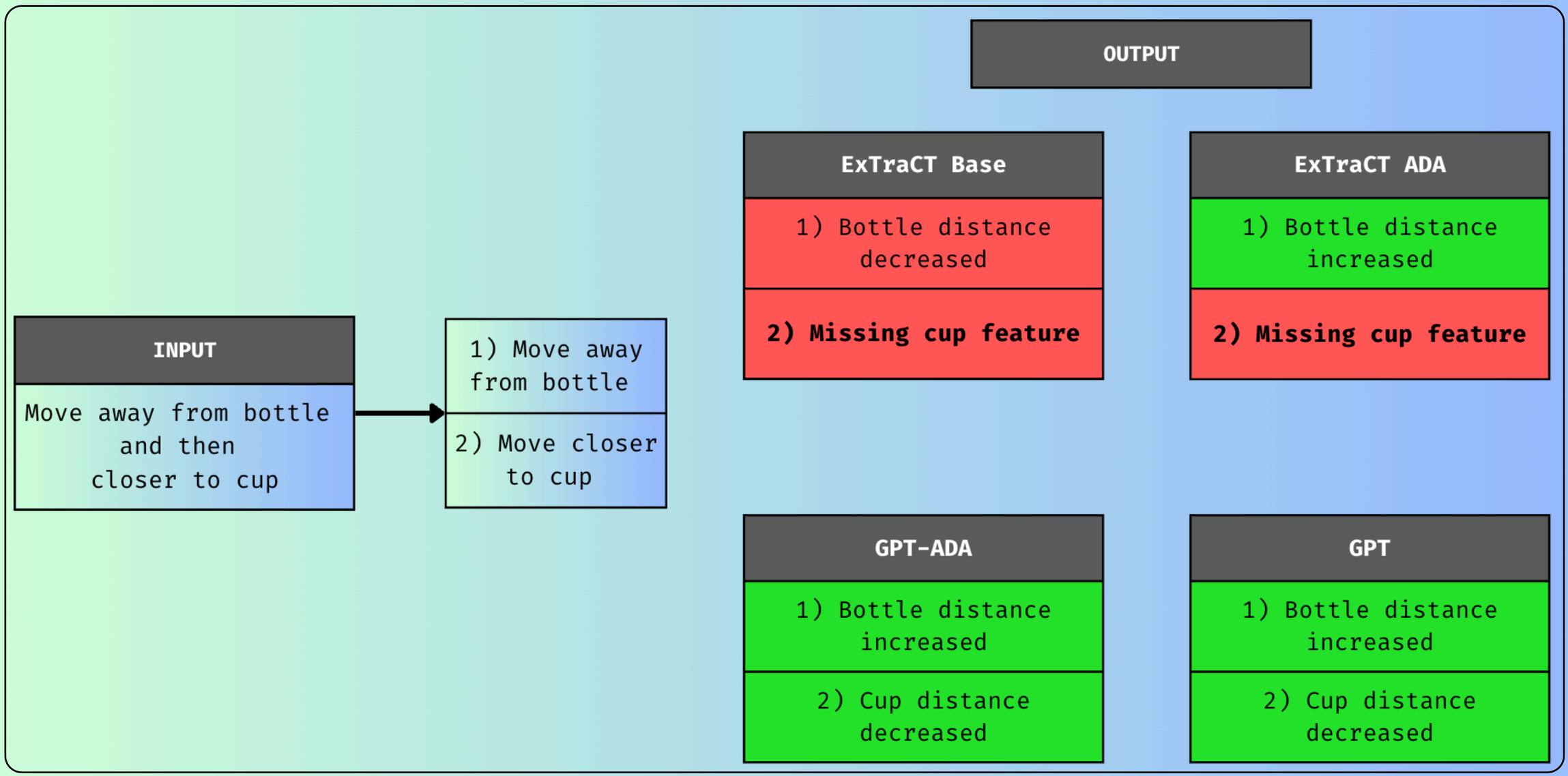
Result

- Performance accuracy for single command
- Performance accuracy for chained commands
- Select best performing approach to replace overall pipeline

Approach	Performance Accuracy		
	Single	Complex (chained)	Total (%)
ExTraCT (Base)	**48/55**	**0/95**	**32.0**
ExTraCT (Ada)	**55/55**	**0/95**	**36.7**
GPT-ADA	**55/55**	**84/95**	**92.6**
GPT	**55/55**	**91/95**	**97.3**



Examples


INPUT
Move away from bottle and then closer to cup
1) Move away from bottle
2) Move closer to cup
OUTPUT
ExTraCT Base
1) Bottle distance decreased
2) Missing cup feature
ExTraCT ADA
1) Bottle distance increased
2) Missing cup feature
GPT-ADA
1) Bottle distance increased
2) Cup distance decreased
GPT
1) Bottle distance increased
2) Cup distance decreased

GPT
VS
GPT-ADA

Observations

- **‘Microwave’/ ‘Bowl’ / ‘Monitor’** leads to feature mismatched for GPT-ADA
- These words perceived as verbs rather than nouns
- GPT-4 has higher success rate

GPT ADA

Input	Move down and closer to the plate and microwave
Split cmds	['Move down', 'closer to the plate', ‘microwave’]
Feature match	[’Z-Cartesian decrease’, ‘plate distance decrease’, ‘microwave distance increase’]
Ground truth	[’Z-Cartesian decrease’, ‘plate distance decrease’, ‘microwave distance decrease’]

GPT

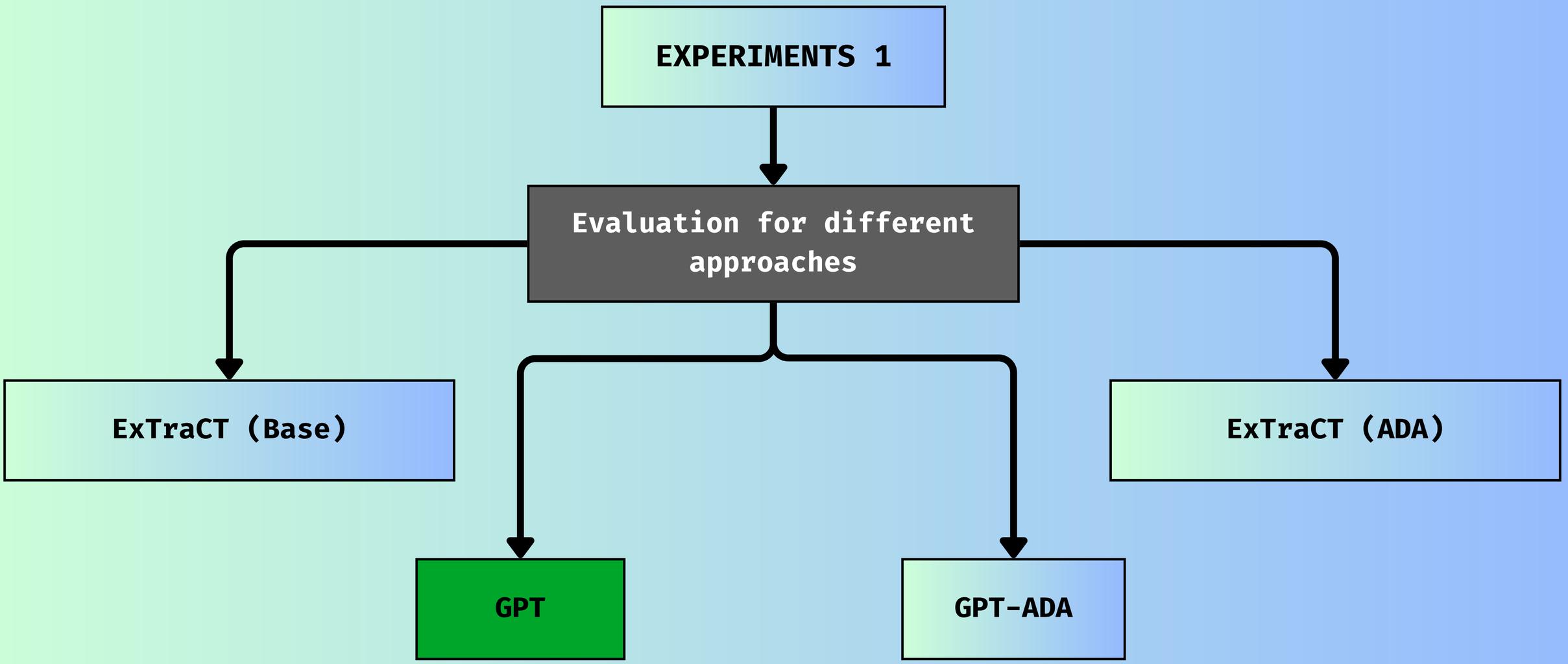
Input	Move down and closer to the plate and microwave
Split cmds	['Move down', 'closer to the plate', ‘microwave’]
Feature match	[’Book distance decrease’, ‘microwave distance decrease’]
Ground truth	[’Z-Cartesian decrease’, ‘plate distance decrease’, ‘microwave distance decrease’]

More Observations

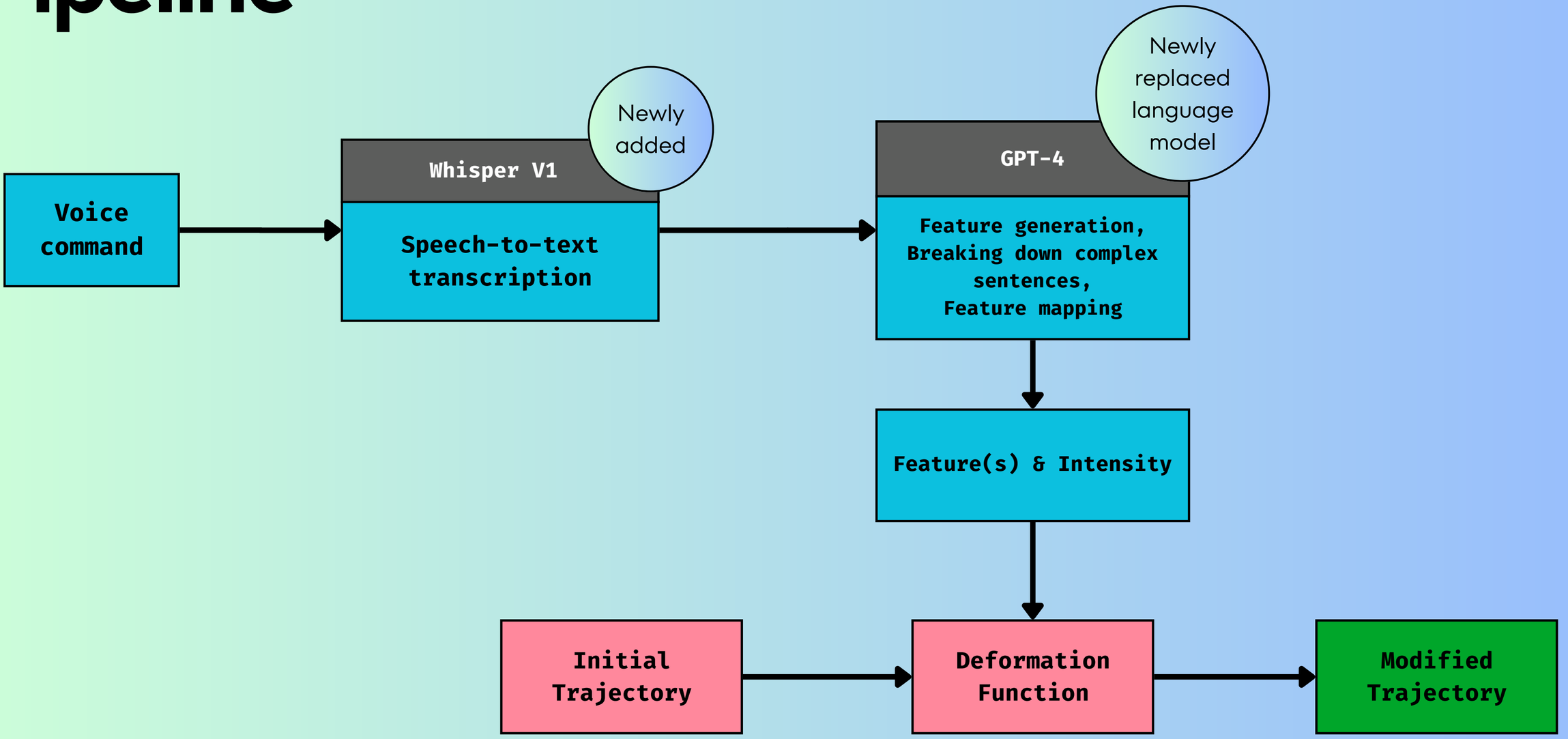
- Adding ‘**the**’ in front of every noun increases success rate

ID	26
Input	Approach the vase and microwave
Split cmds	['Approach the vase, 'microwave']
Feature match	[‘vase distance decrease’, ‘microwave distance increase’]
Ground truth	[‘Cup distance decrease’, ‘microwave distance decrease’]

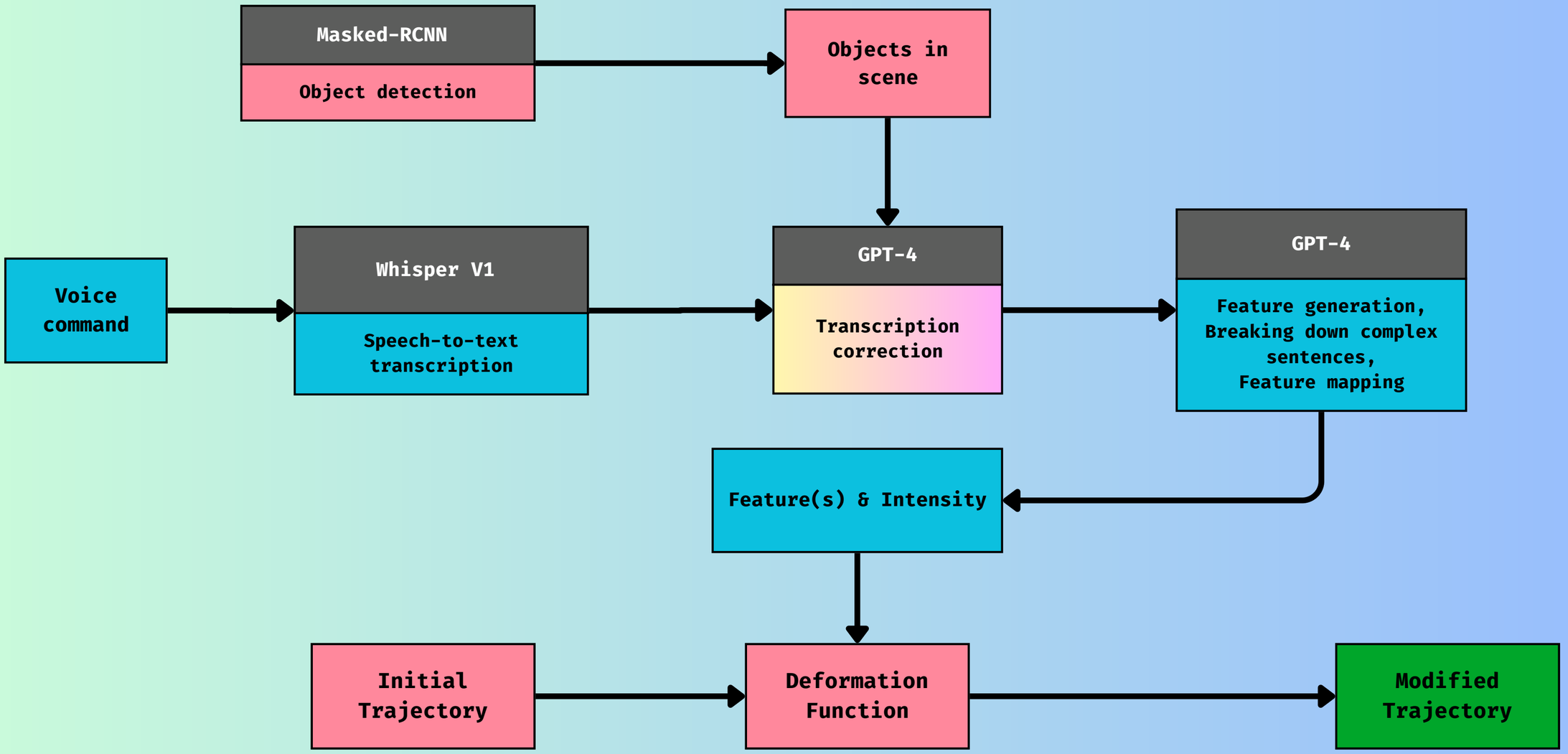
ID	27
Input	Approach the glass and the microwave
Split cmds	['Approach the glass, ‘Approach the microwave’]
Feature match	[‘glass distance decrease’, ‘microwave distance decrease’]
Ground truth	[‘Cup distance decrease’, ‘microwave distance decrease’]


EXPERIMENTS 1
Evaluation for different approaches
ExTraCT (Base)
GPT
GPT-ADA
ExTraCT (ADA)

Non-optimised Pipeline



Optimised Pipeline



Challenges of Transcription

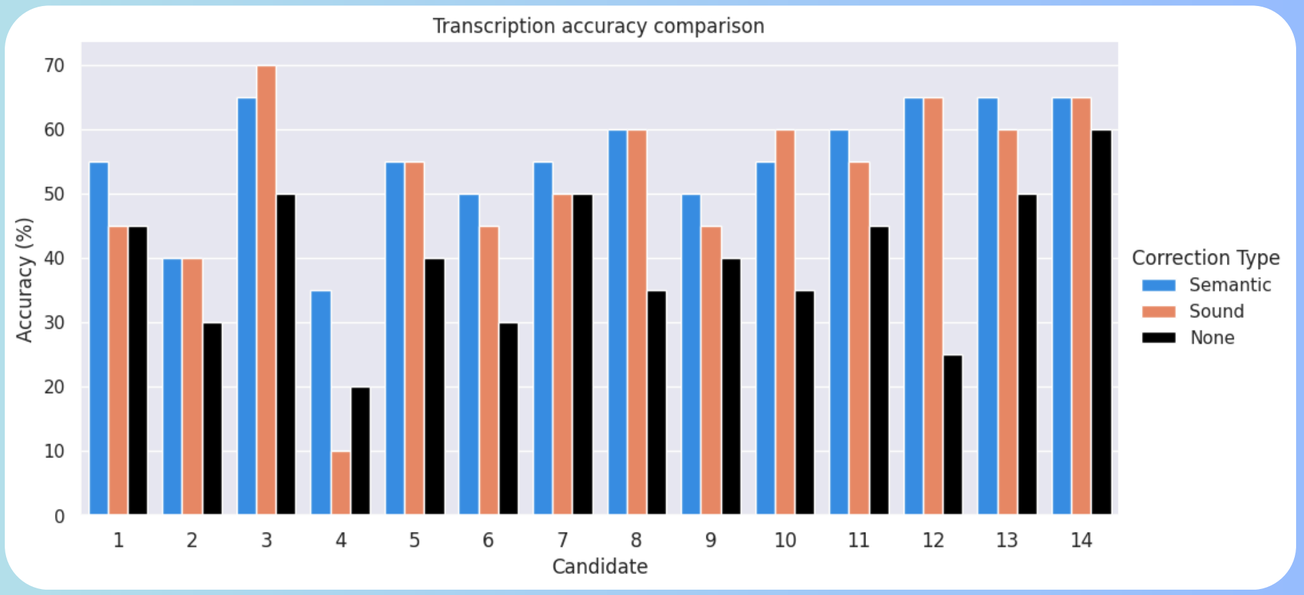
Challenges	Examples	Action taken
Object names can have multiple synonyms	• Approach the cup • Approach the mug	• Utilise scene objects as a layer of check (ground truth) • Add another API call to rectify transcription of object
Different noun objects may have similar pronunciations	• Approach the mark • Approach the mug	

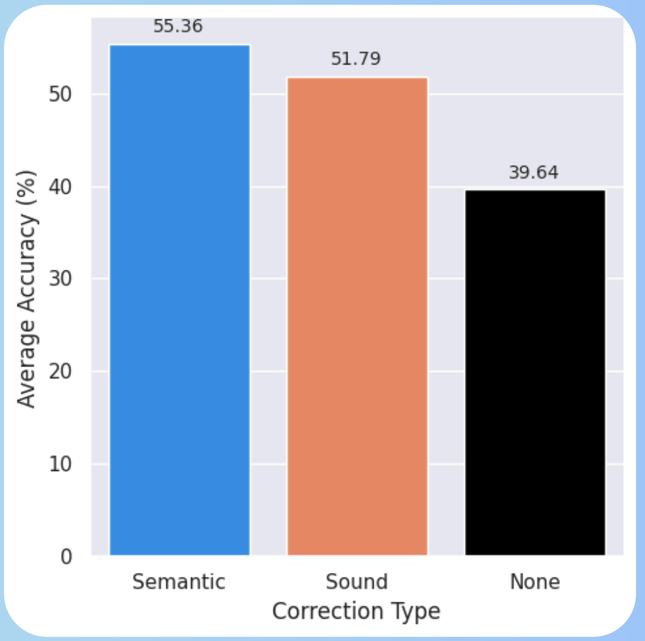
Experiment 2

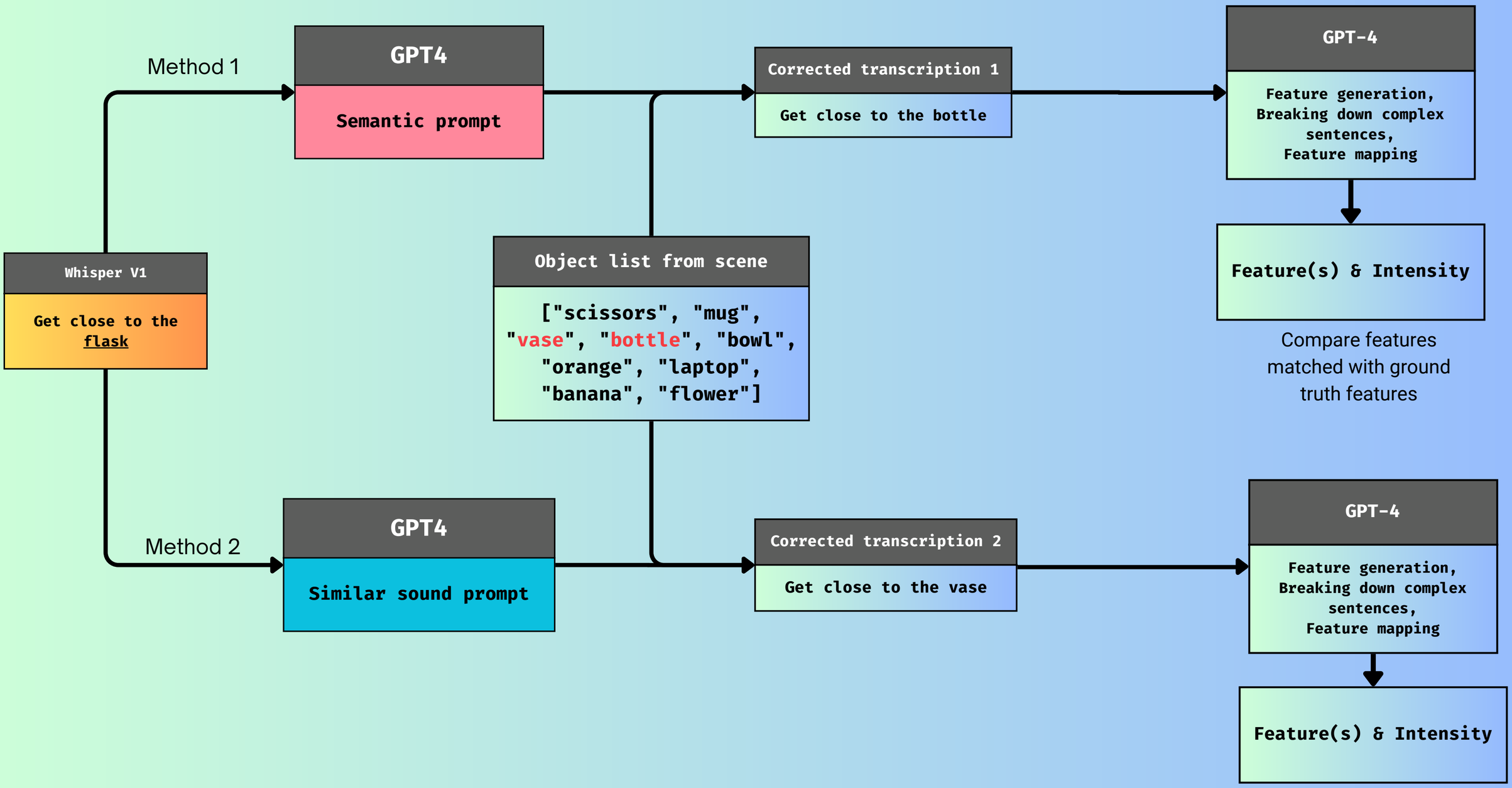
- Assess effectiveness of Whisper in transcribing local accent
- Semantic & similar sound prompts tested separately for transcription correction
- Audio data sampled across 14 candidates of different genders & races
- 10 data with 2 iterations each

Result

METHOD	Accuracy in Feature Matching (%)
No Rectification	39.6
Semantic	55.4
Similar Sound	51.8






Method 1
GPT4
Semantic prompt
Corrected transcription 1
Get close to the bottle
GPT-4
Feature generation,
Breaking down complex
sentences,
Feature mapping
Feature(s) & Intensity
Whisper V1
Get close to the
flask
Object list from scene
["scissors", "mug",
"vase", "bottle", "bowl",
"orange", "laptop",
"banana", "flower"]
Compare features
matched with ground
truth features
Method 2
GPT4
Similar sound prompt
Corrected transcription 2
Get close to the vase
GPT-4
Feature generation,
Breaking down complex
sentences,
Feature mapping
Feature(s) & Intensity

Similar sound > Semantic meaning

Object List	["Scissors", "mug", "vase", "bottle", "bowl", "orange", " laptop", "banana", "flower"]

Prompt	Similar Sound
Input	Approach the mark and keep away from the bottle
Output	Approach the mug and keep away from bottle
Reasoning	"Mug" is the closest sounding sound from the list to "mark" due to the ending sound ('-ug' and '-ark')

Prompt	Semantic Meaning
Input	Approach the mark and keep away from the bottle
Output	Approach the flower and keep away from bottle
Reasoning	"Flower" might be a suitable replacement if we think of "mark" in a decorative context

Pass

Fail

Similar sound < Semantic meaning

Object List	["Scissors", "mug", "vase", "bottle", "bowl", "orange", " laptop", "banana", "flower"]

Prompt	Similar Sound
Input	Move a bit closer to bowl but stay away from flask
Output	Move a bit closer to bowl but stay away from flower
Reasoning	"Flower" phonetically resembles "Flask"

Fail

Prompt	Semantic Meaning
Input	Move a bit closer to bowl but stay away from flask
Output	Move a bit closer to bowl but stay away from bottle
Reasoning	Closest meaning to "flask" is "bottle"

Pass

Observations	
1	Optimal transcription correction method is context dependent
2	Reducing the number of objects in scene improves accuracy
3	Whisper does not work well for local accent

Future Works

Use Whisper V2 model

Use real-time speech-to-text model for faster inference

Fine-tuning Whisper with local accent

Demo

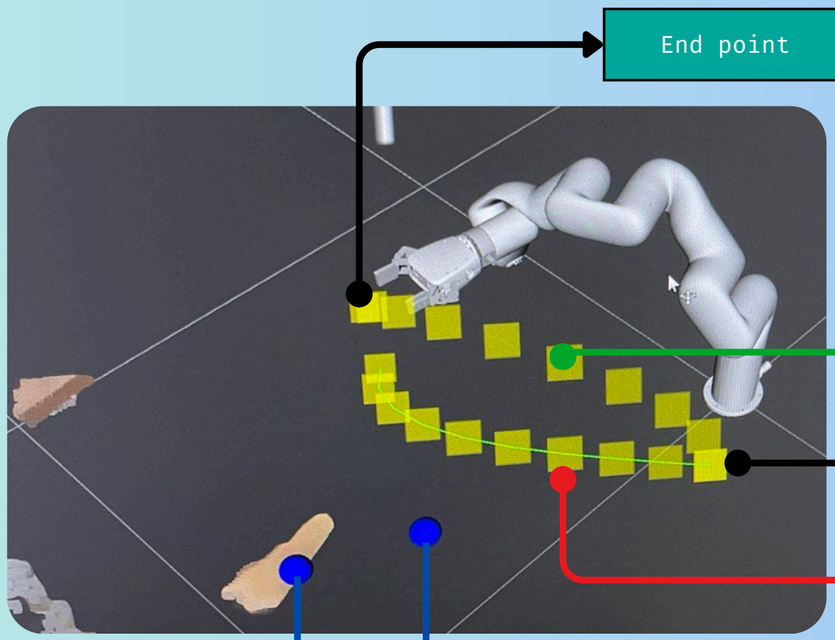
Command	Stay away from the bottle

Initial Trajectory

Modified Trajectory (Neutral intensity)

Demo Trajectory

Voice command
Stay away from the bottle



Conclusion

- Improved performance for feature classification
- Increased robustness for consecutive chained commands
- Incorporated speech-to-text capability
- Integrated overall pipeline with robotic arms

Thank You

References

[1] A. Bucker, L. F. C. Figueredo, S. Haddadinl, A. Kapoor, S. Ma, and R. Bonatti, “Reshaping robot trajectories using Natural Language Commands: A study of Multi-Modal Data alignment using Transformers,” 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Oct. 2022, doi: 10.1109/iros47612.2022.9981810.

[2] A. Bucker et al., “LATTE: LAnguage Trajectory TransformEr,” Arxiv, May 2023, doi: 10.1109/icra48891.2023.10161068.

[3] J. L. Yow, N. Garg, M. Ramanathan, and W. T. Ang, “ExTraCT -- Explainable Trajectory Corrections from language inputs using Textual description of features,” arXiv (Cornell University), Jan. 2024, doi: 10.48550/arxiv.2401.03701.

[4] “GPT-4,” OpenAI. https://openai.com/index/gpt-4-research

Q&A