

---

# Persona-aware Multi-Party Response Generation

Khyati Mahajan 

Samira Shaikh

 [kmahaja2@uncc.edu](mailto:kmahaja2@uncc.edu)

---

# Agenda



---

# MPC Resources

- Open domain, unscripted written corpora are mostly scraped from social networks like Reddit, Twitter, Ubuntu - restrictions on scraping data, changing TOS
- None of the existing corpora provide persona information for each speaker and addressee

---

# Persona-aware MPC response generation

- Persona-aware generation has been shown to generate more consistent and engaging responses in 2PC
- RG research limited to modeling few corpora
  - Mostly Ubuntu IRC
- Only one persona-aware response generation model for multi-party conversation modeling
  - Not publicly available
  - Models TV scripts - not real world data

---

# Data collection from user experiments

- Collected **>2500 conversations** across 3 user studies
  - ~30k utterances
  - Focused on open domain MPC
  - Includes political discussions on specific events
- Data available upon request
- We utilize this dataset towards persona-aware response generation

# Data collection statistics

	Exp 1	Exp 2	Exp 3
<b>Time Period</b>	April 2021	October 2021	March-April 2022
<b>Race</b>	80% white and 20% non-white	77% white and 23% non-white	81% white and 19% non-white
<b>Gender</b>	50% female, 49% male and 1% other	57% female, 42% male and 1% other	52% female, 47% male and 1% other
<b>Leaning</b>	51.5% liberal, 42.5% conservative, and 6% independent	42% liberal, 41% conservative, 17% independent	51% liberal, 44% conservative and 5% independent

---

# Persona attributes for user behavior

- Collected attributes via user surveys during participant recruitment
  - Race
  - Gender
  - Leaning
- Generated annotations for user behavior
  - Fall into one of 4 categories
  - Spectators, Expressors, Suppressors, Avoiders

# User behaviors on social media

Spectators	Expressors	Suppressors	Avoiders
prefer to observe emotional conversations unfolding	utilize social media as a place to obtain information from or a place to keep in contact with family and friends	suppress overly emotional content on social media	discerning and cautious in their emotion sharing
discerning and cautious in their emotion sharing	find the spread of emotions to be a positive goal in and of itself - consider social media similar to real life for discourse	actively engage in discourse with Expressors by attempting to advance facts and advocate for suppressing the emotion expression	prefer to discuss difficult topics but mainly share content they find positive, unifying, or productive

---

# Generating user behavior annotations

- Zero-shot prompting strategy with `flan-t5-xxl`

```
Instruction: Classify User1 into one of the 4  
categories as defined below:
```

```
Spectators: <definition>
```

```
Expressors: <definition>
```

```
Avoiders: <definition>
```

```
Suppressors: <definition>
```

```
User1: I still don't think we should have to have proof  
of vaccinations to go anywhere, when masks were supposed  
to be working all along.
```

```
User2: You already need proof of several other  
vaccinations in order to attend school, go abroad, or  
work in certain fields.
```

```
User1: That is true but this is slightly different,  
it's too new for some
```

---

# Generating user behavior annotations

- We find that the model performs more deterministically when the users are explicitly mentioned in the prompt
- On average, **70.1% annotations reflect the user behavior well**, **29.9% annotations are modified to reflect user behavior better**

# User behavior annotations stats

Exp No.	Total Users	Annotated by	Behaviors			
			Avoiders	Expressors	Spectators	Suppressors
1	121	flan-t5-xxl	7	62	41	11
		Manually corrected	18	76	19	8
2	140	flan-t5-xxl	7	70	46	17
		Manually corrected	12	91	23	14
3	182	flan-t5-xxl	10	102	66	4
		Manually corrected	23	111	38	10

- Expressors form the clear majority of behaviors in our experiments, whereas Suppressors are fewer in number
- Most users classified as Avoiders did not post much during the entire experiment, whereas those classified as Spectators preferred engaging with non-political content

---

## Final formatted dataset stats

---

Statistic	Exp 1	Exp 2	Exp 3
Conversations with >5 utterances	550	563	720
Total no of turns	6384	5242	9845
Avg turns per conversation	11.61	9.31	13.67
Total no of tokens	97142	83995	144615
Avg tokens per turn	15.21	16.02	14.69
Avg tokens per conversation	15.71	15.55	14.34
Vocab size	9039	8520	11047
Total users	122	144	187
Avg users in conversation	6.55	6.75	9.41

---

Train - 1766 conversations, Valid - 313 conversations, Test - 521 conversations

---

---

## Existing work in MPCRG

---

Paper	Architecture	Description
GSN	LSTM/GRU	Introduces graphs to model conversational structure
MPC-BERT	BERT	Tracks both speaker, addressee info
HeterMPC	GSN + Transformer	Heterogeneous graph modeling interlocutor and utterance information
PersonaTKG	GRU + Attention	Graph modeling utterance and persona information

---

---

---

# Task Description

- Given conversation history with each data point representing (*speakers, addressees, personas, history*), generate corresponding (*response*)
- Model conversation structure
- Utilize persona-based information towards more consistent response generation for each participant

---

# Task Description

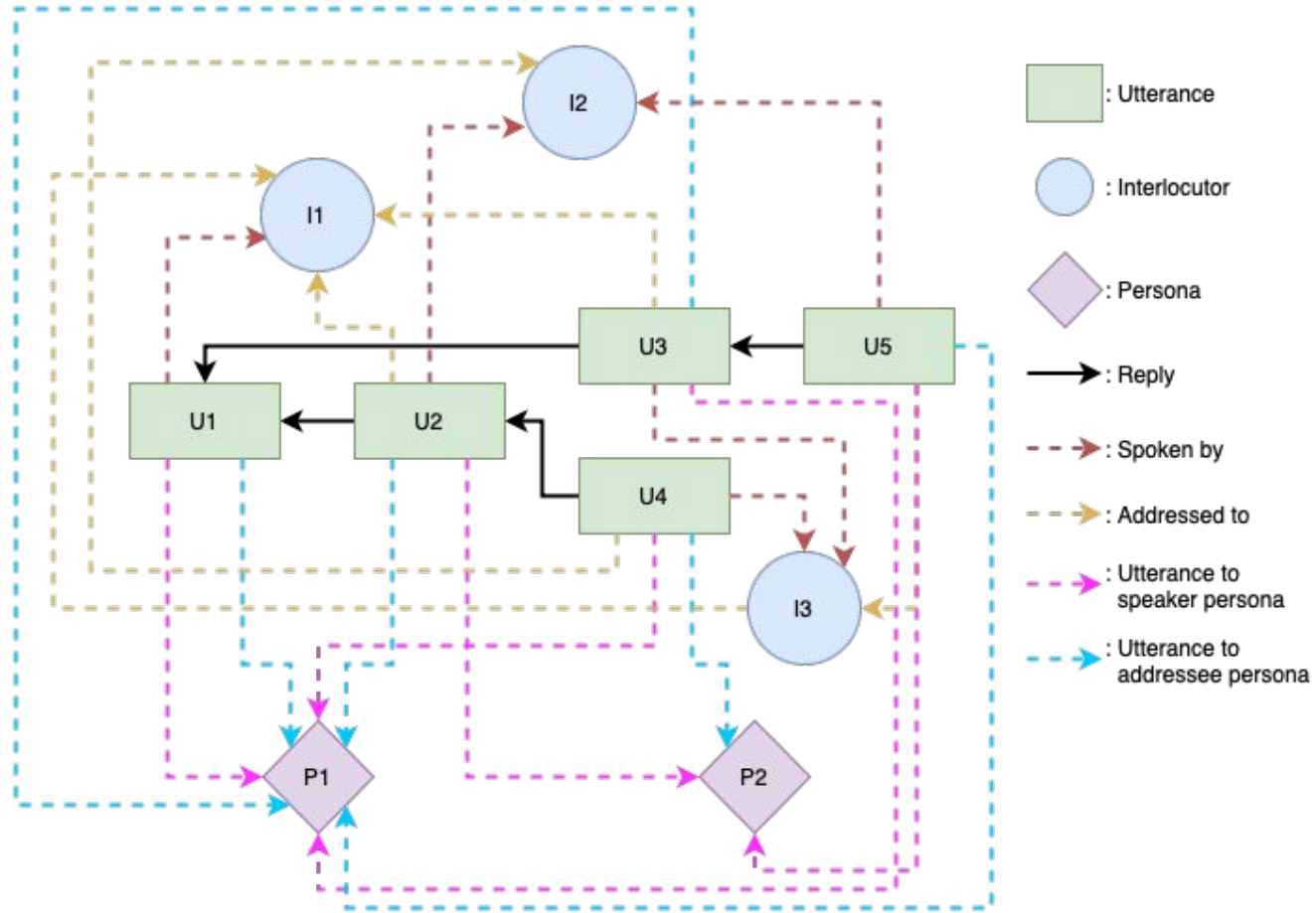
- $G$  is a heterogeneous graph modeling utterance context and interlocutor relations to utterances
- $r_k$  and  $r_{<k}$  stand for the  $k$ -th token and the first  $(k - 1)$  tokens of response  $r$  respectively.  $|r|$  is the length of  $r$ .

$$\begin{aligned}\bar{r} &= \operatorname{argmax}_r \log P(r|G) \\ &= \operatorname{argmax}_r \sum_{k=1}^{|r|} \log P(r_k | G r_{<k}).\end{aligned}$$

---

# Conversation Graph Modeling

- Modeling conversation structure with graph-based Transformer networks
- Nodes encode information for
  - Utterance (*+persona with approach<sub>1</sub>*)
  - Interlocutor
  - (*Persona with approach<sub>2</sub>*)
- Edges encode relationships between nodes
  - 6 kinds (*+4 kinds for persona with approach<sub>2</sub>*)



Conversation structure in multi-party conversation modeling (shown with unidirectional edges for brevity)

---

# Base Architecture

- Relationships between utterances and their corresponding speakers and addressees are modeled with *heterogeneous graphs* using the Deep Graph Library<sup>[46]</sup>
  - Node ( $\mathbb{N}$ ) types and edge ( $\mathbb{E}$ ) types are separately modeled
- Speakers and addressees are encoded in speaking order for each conversation
- Utterance nodes are encoded as beginning of sequence tokens

---

## Base Architecture (contd)

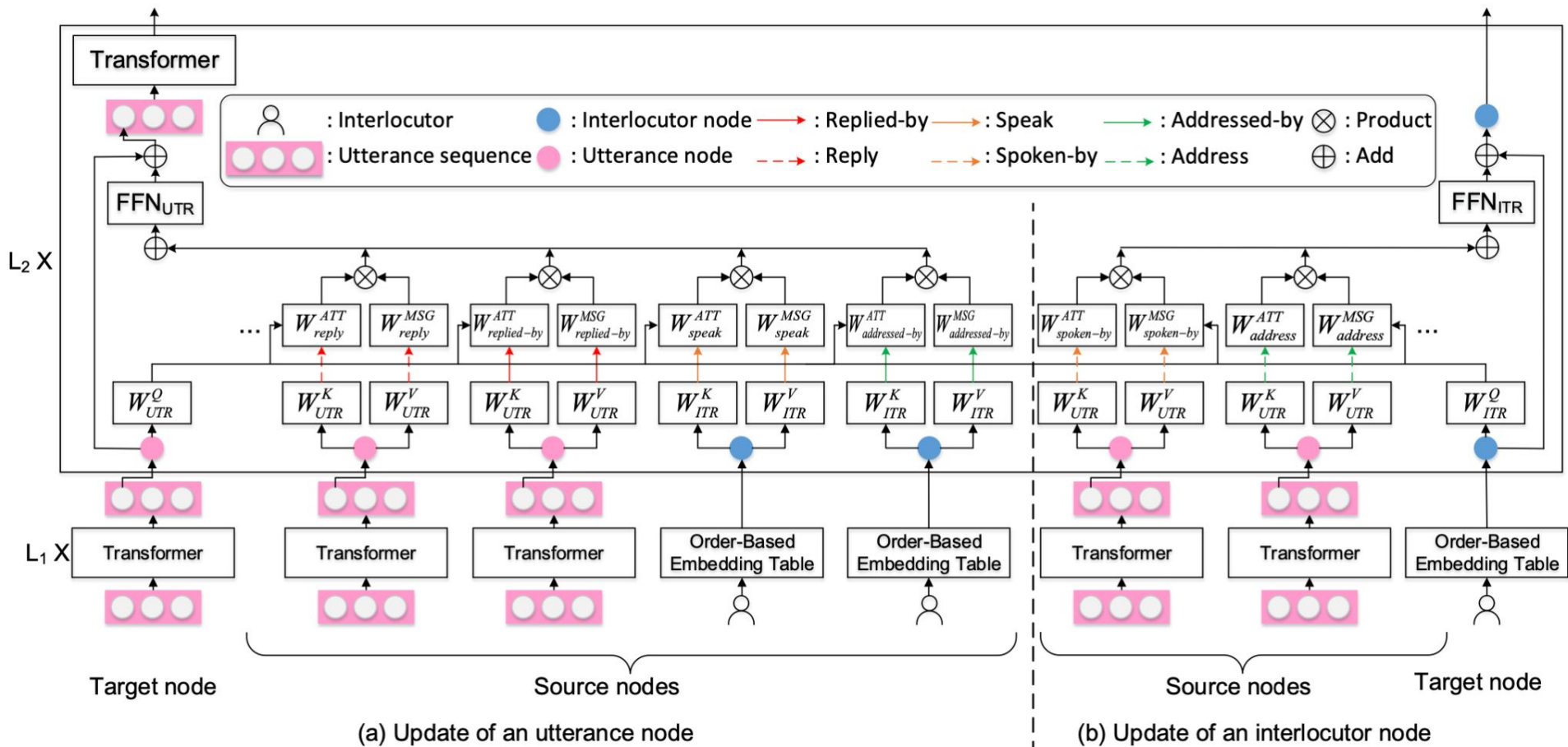
- Utterance content is encoded via `BERT-base-uncased`
- Updates for complete utterance encodings are handled with an additional Transformer layer
- Representations of an utterance node before ( $\mathbf{h}_t^l$ ) and after node updating ( $\mathbf{h}_t^{l+1}$ ) are concatenated and then compressed by a linear transformation ( $\hat{\mathbf{h}}_t^{l+1}$ )

$$\hat{\mathbf{h}}_t^{l+1} = [\mathbf{h}_t^l; \mathbf{h}_t^{l+1}] \mathbf{W}_{com} + \mathbf{b}_{com},$$

---

## Base Architecture (contd)

- Representations are updated by feeding them into the graph for absorbing context information
- *Heterogeneous attention* weights are calculated between connected nodes and pass *messages over the graph* in a N-E-type-dependent manner
  - Using parameters to maximize feature distribution differences for modeling heterogeneity
  - N-type-dependent FFN followed by a residual connection to *aggregate information*



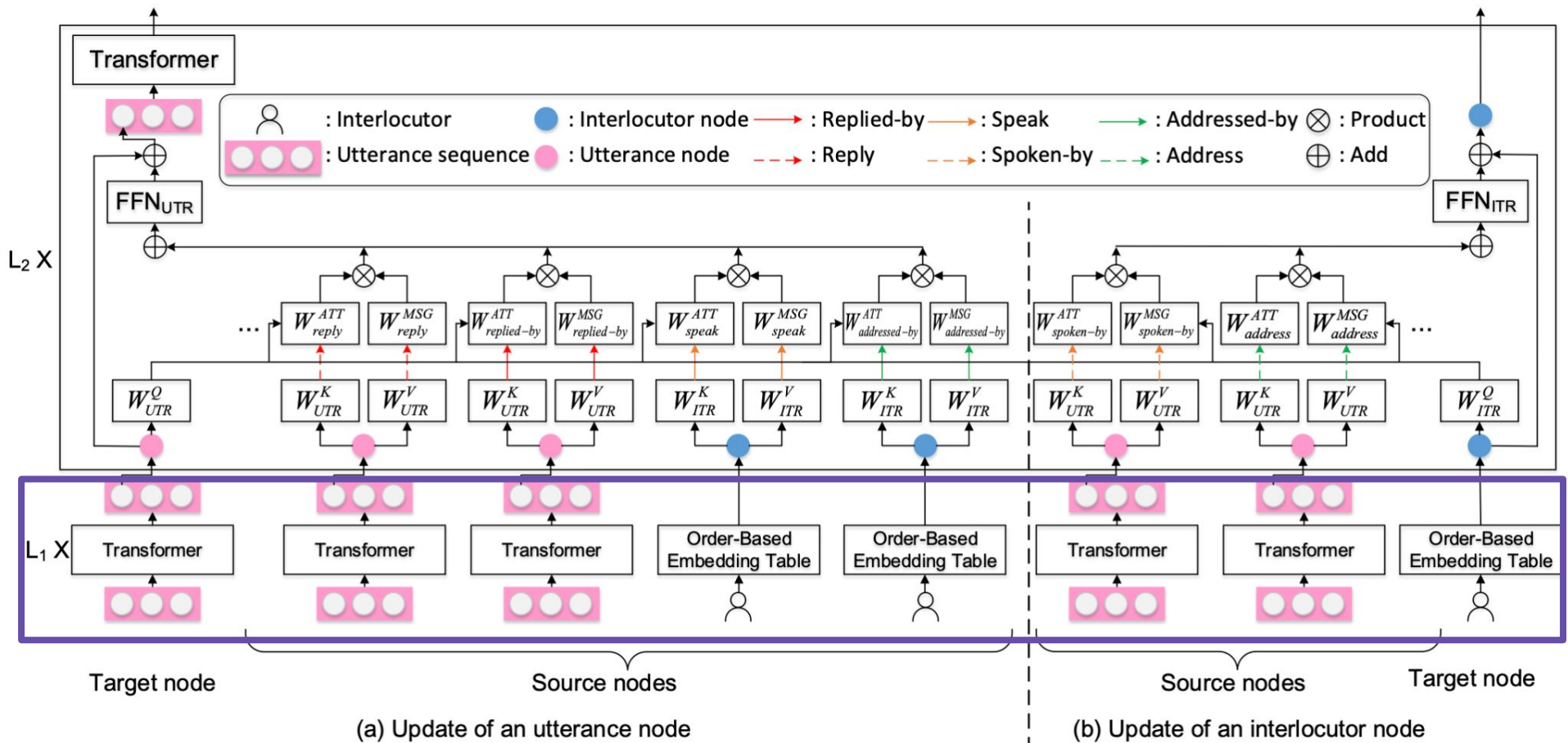
---

## Node Initialization

- *Utterances*: [CLS] and [SEP] marking BOS and EOS, encoded with a Transformer layer

$$\mathbf{H}_m^{l+1} = \text{TransformerEncoder}(\mathbf{H}_m^l)$$

- *Interlocutors*: embedding vector derived by looking up an order-based interlocutor embedding table



---

# Heterogeneous Attention

- N-type-dependent linear transformations are applied to node representations before attention calculation to ensure heterogeneous nodes share similar feature distributions
- Each  $\mathbb{E}$  type is assigned a separate linear projection so that the semantic relationship between two connected nodes can be accurately described when calculating attention weights

# Heterogeneous Attention (contd)

- Representations of the source and target  $\mathbb{N}$  at the  $l$ -th iteration  $(h_s^l, h_t^l)$ , serve as *key* ( $\mathbb{K}$ ) and *query* ( $\mathbb{Q}$ ) vectors of attention calculation towards heterogeneous attention weight  $(w^l(s, e, t))$

$$\mathbf{k}^l(s) = \mathbf{h}_s^l \mathbf{W}_{\tau(s)}^{\mathbb{K}} + \mathbf{b}_{\tau(s)}^{\mathbb{K}},$$

$$\mathbf{q}^l(t) = \mathbf{h}_t^l \mathbf{W}_{\tau(t)}^{\mathbb{Q}} + \mathbf{b}_{\tau(t)}^{\mathbb{Q}},$$

$$w^l(s, e, t) = \mathbf{k}^l(s) \mathbf{W}_{e_s, t}^{ATT} \mathbf{q}^l(t)^T \frac{\mu_{e_s, t}}{\sqrt{d}}.$$

Edge-type dependent  
linear projection

Adaptive factor  
scaling to attention

Embedding  
vector dim

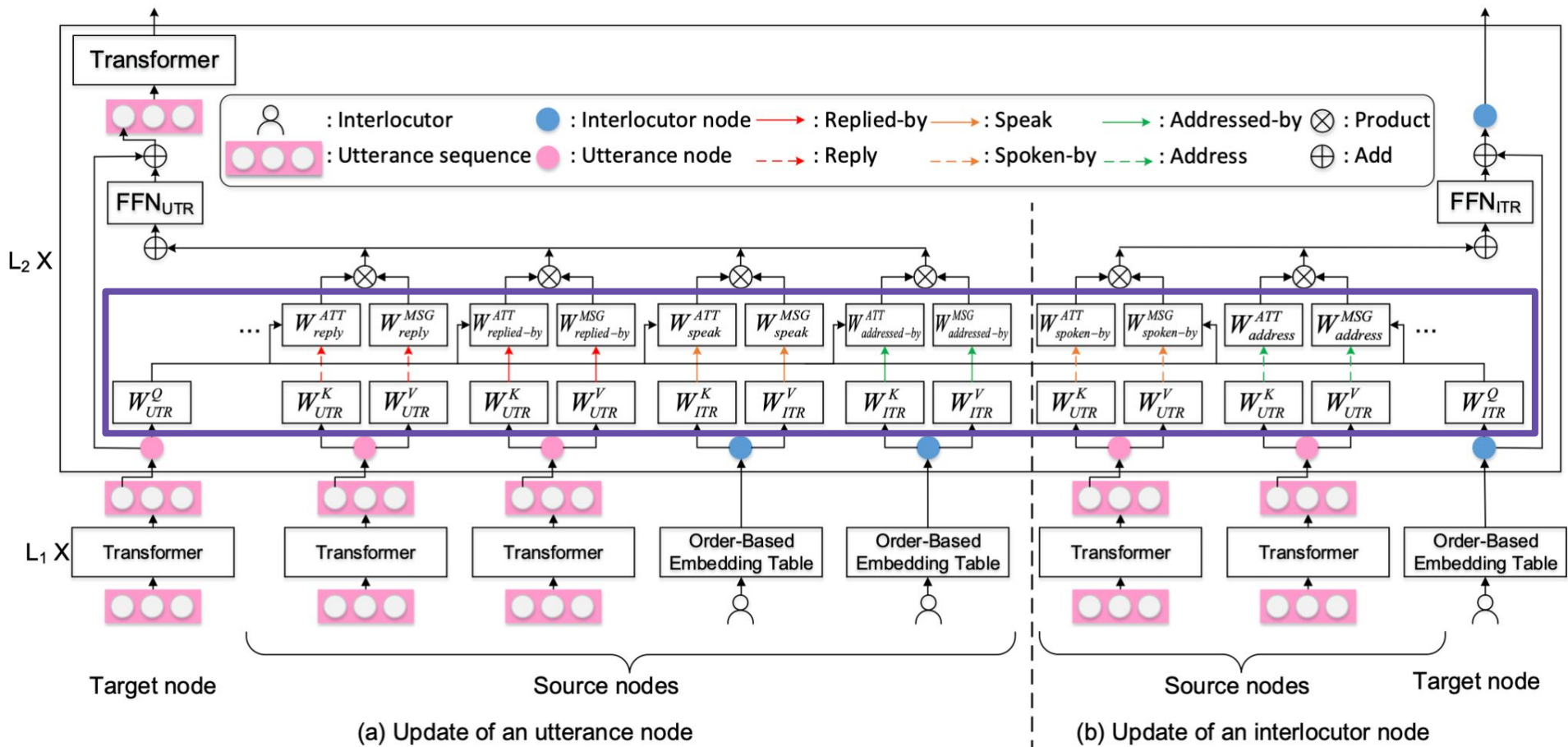
---

# Heterogeneous Message Passing

- When passing the message of a source  $\mathbb{N}$  that serves as a *value* ( $V$ ) vector to a target  $\mathbb{N}$ ,  $\mathbb{N}$ - $\mathbb{E}$ -type-dependent parameters are introduced considering the heterogeneous properties of nodes and edges

$$\mathbf{v}^l(s) = \mathbf{h}_s^l \mathbf{W}_{\tau(s)}^V + \mathbf{b}_{\tau(s)}^V,$$

Passed message  $\longrightarrow \bar{\mathbf{v}}^l(s) = \mathbf{v}^l(s) \mathbf{W}_{e_{s,t}}^{MSG}$



---

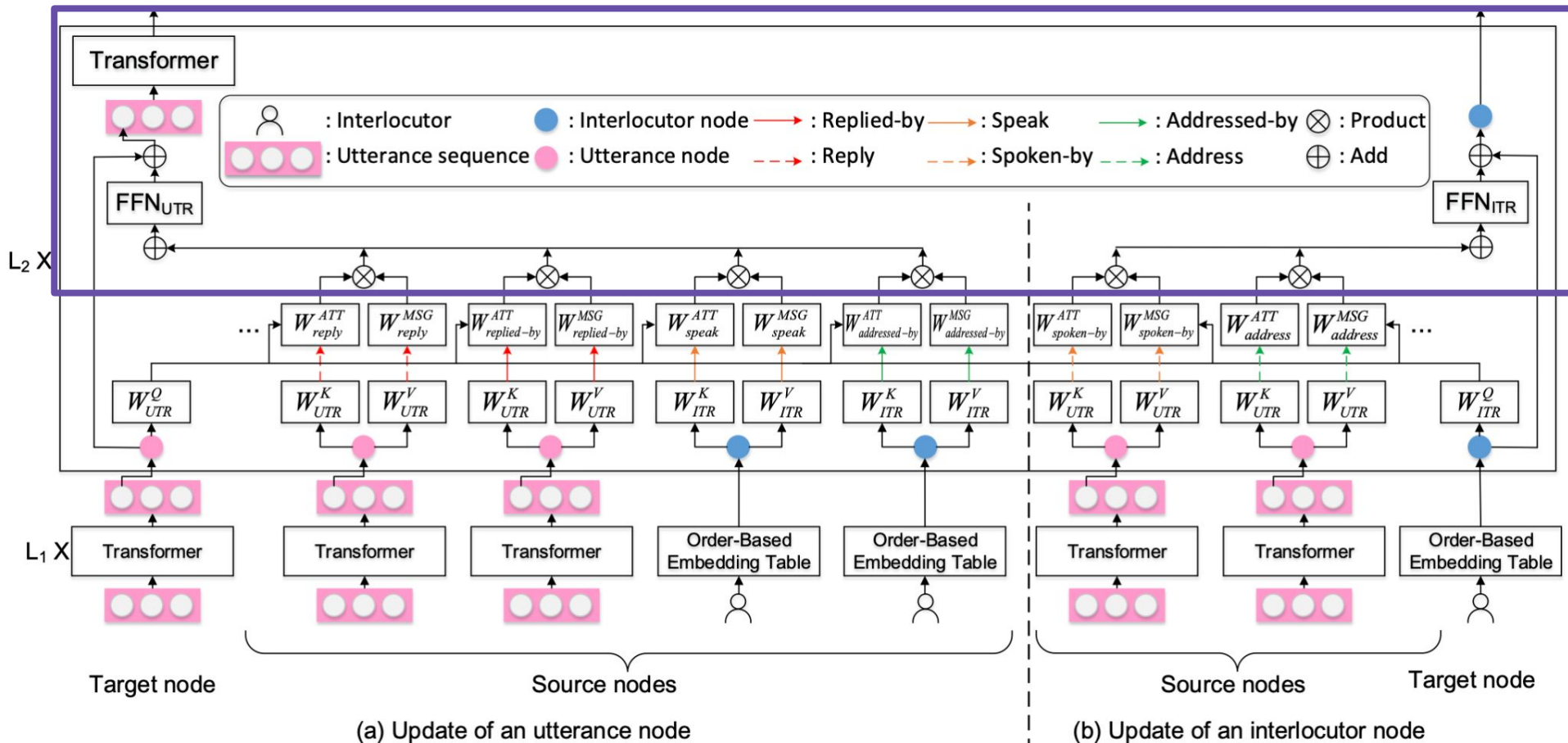
# Heterogeneous Aggregation

- Softmax to normalize attention weights + summarize messages from all source nodes

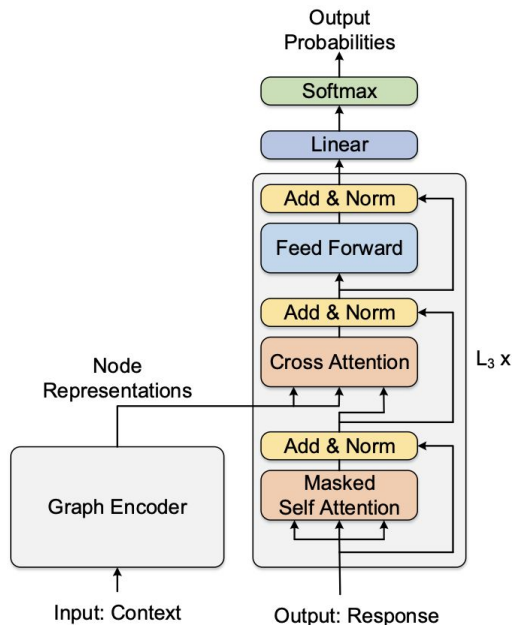
$$\bar{\mathbf{h}}_t^l = \sum_{s \in \mathcal{S}(t)} \text{softmax}(w^l(s, e, t)) \bar{\mathbf{v}}^l(s),$$

- Then aggregate with original node representation using a N-type-dependent FFN followed by a residual connection

$$\mathbf{h}_t^{l+1} = \text{FFN}_{\tau(t)}(\bar{\mathbf{h}}_t^l) + \mathbf{h}_t^l,$$



# Decoder



- Masked self-attention operation is first performed so each token cannot attend to future tokens - avoids information leakage
- Cross-attention operation over the node representations of the graph encoder output is performed to incorporate graph information for decoding

---

# Approach 1 - PersonaHeterMPC<sub>concat</sub>

- RQ: How does modeling personas as text along with utterance encodings perform towards persona-aware RG?
- Speaker and addressee personas are provided as inputs, with [CLS] and [SEP] marking BOS and EOS
  - Encoder input changes to  $H = \{ (p_{u1}^{spk}, p_{u1}^{adr}, h_{u1}) , \dots \}$
  - Decoder input to  $D = \{ p_{ans}^{spk}, p_{ans}^{adr}, [BOS] \}$
- Computation for loss is updated by masking persona positions

---

# Approach 1 - more experiments

- We try 3 strategies
  1. Laconic vs descriptive persona

White male conservative expressor

VS

I am a white male with a conservative ideology. I usually prioritize emotional expression on social media, and view it as a platform to share powerful and important content.

Descriptive persona is generated based on template:

I am a [race] [gender] with a [leaning] ideology. [user behavior definition].

---

# Approach 1 - more experiments

- We try 3 strategies
  1. Laconic vs descriptive persona
  2. Using different special tokens such as [BOSP], [EOSP], [BOAP] and [EOAP] to demarcate the persona sequences

---

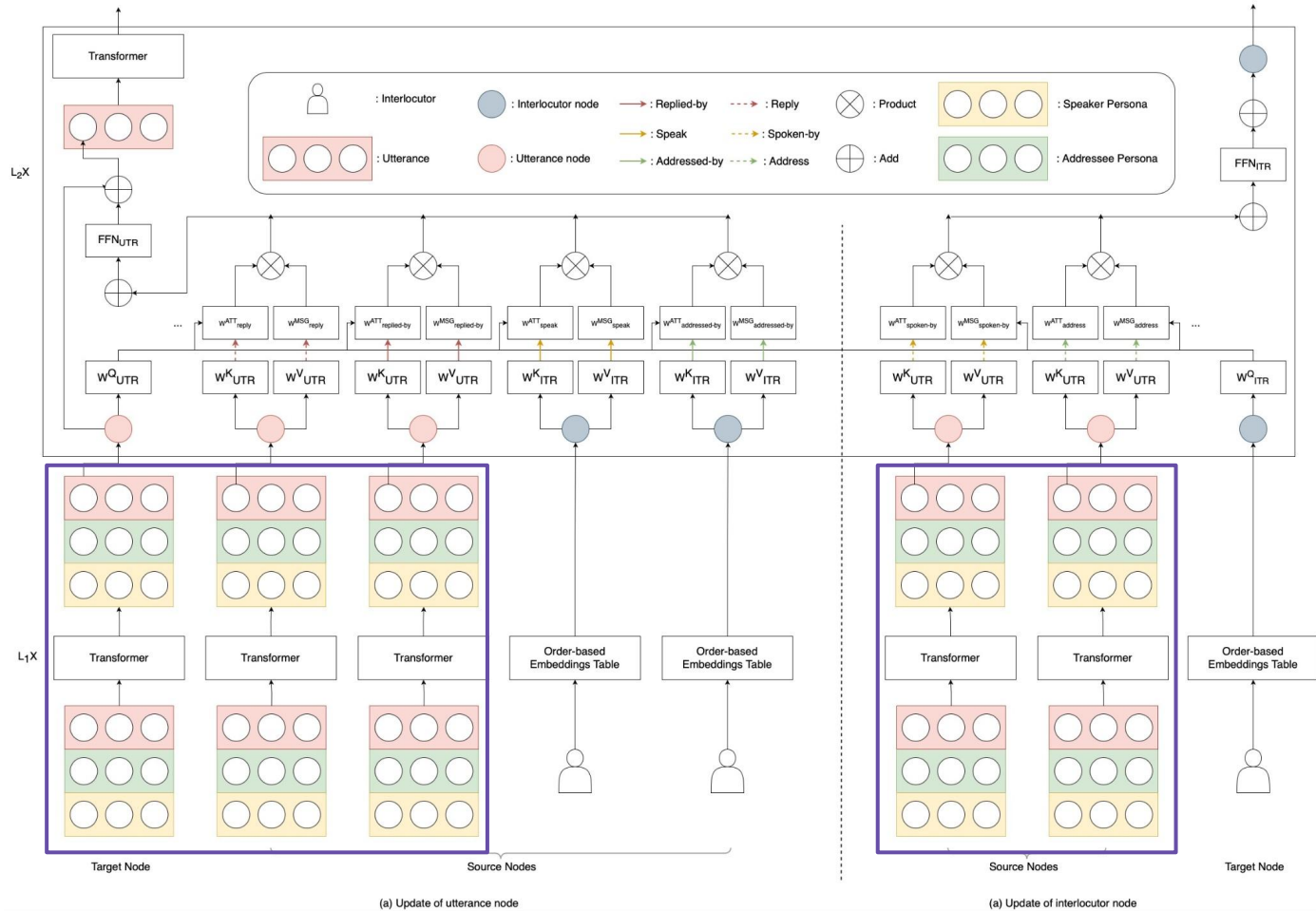
# Approach 1 - more experiments

- We try 3 strategies
  1. Laconic vs descriptive persona
  2. Using different special tokens such as [BOSP], [EOSP], [BOAP] and [EOAP] to demarcate the persona sequences
  3. Oversampling the dataset by a factor of 5

---

# Approach 1 - more experiments

- We try 3 strategies
  1. Laconic vs descriptive persona
  2. Using different special tokens such as [BOSP], [EOSP], [BOAP] and [EOAP] to demarcate the persona sequences
  3. Oversampling the dataset by a factor of 5
- We find that
  - Descriptive personas perform better
  - Special tokens and oversampling does not help

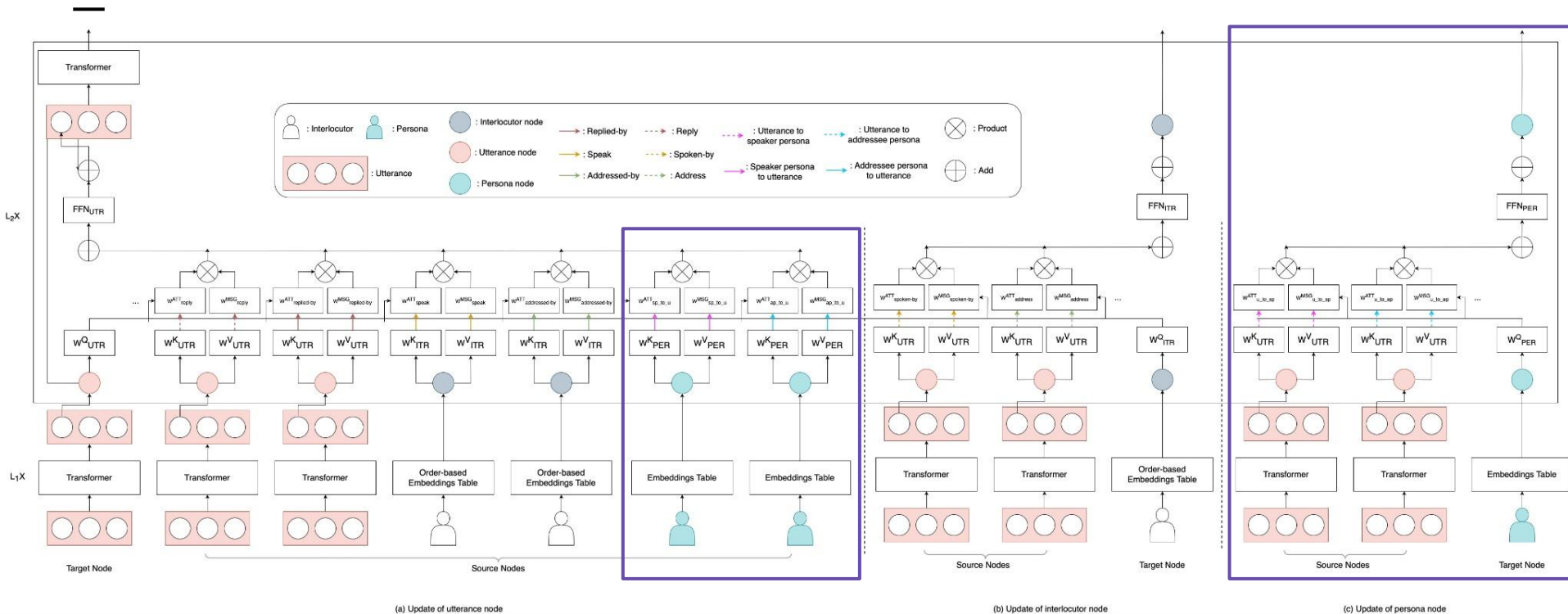


# PersonaHeterMPC<sub>concat</sub> with persona + utterance encodings

---

## Approach 2 - PersonaHeterMPC<sub>graph</sub>

- RQ: How does modeling personas as graph nodes connected to utterance nodes perform towards persona-aware RG?
- We introduce
  - New node type - persona
  - New edges - `utt-to-spk-persona`, `spk-persona-to-utt`, `utt-to-adr-persona`, `adr-persona-to-utt`
- Persona nodes are indexed via global lookup table, created from laconic persona types in the dataset



PersonaHeterMPC<sub>graph</sub> with persona nodes connected to utterance nodes via four edge types

---

# Model Evaluation

- Automatic metrics
  - Context-free: BLEU, ROUGE, METEOR
- Human evaluation
  - Fluency
  - Relevance (content)
  - Informativeness
- To compare performance with HeterMPC

---

# Human Evaluation MPC Metrics

- Extend human evaluation to gauge task-specific performance towards engaging and consistent responses
  - *Initiative-taking* - Was initiative taken by the response? Did it help move the conversation forward?
  - *Context relevance* - Does response makes sense as part of the conversation thread at the graph node?
  - *Persona relevance* - Suitability considering the speaker and addressee personas in regards to user behavior

# Model Performance - Automatic

Model	(top_p, top_k)	Metrics					
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE <sub>L</sub>
HMPC	(0.9, 5)	12.091	4.967	2.558	1.701	5.076	9.377
PHMPC <sub>concat</sub>	(0.9, 5)	<b>13.118</b>	4.740	2.066	1.121	4.960	6.979
PHMPC <sub>graph</sub>	(0.9, 5)	12.784	<b>5.834</b>	<b>3.697</b>	<b>2.859</b>	<b>5.338</b>	9.013
HMPC	(0.5, 5)	11.712	4.894	2.940	2.244	4.978	<b>9.612</b>
PHMPC <sub>concat</sub>	(0.5, 5)	11.305	4.358	2.068	1.285	4.594	6.574
PHMPC <sub>graph</sub>	(0.5, 5)	<b>12.367</b>	<b>5.643</b>	<b>3.652</b>	<b>2.902</b>	<b>5.153</b>	9.020
HMPC	(0.9, 10)	11.747	4.696	2.727	1.993	4.869	<b>8.263</b>
PHMPC <sub>concat</sub>	(0.9, 10)	<b>12.085</b>	4.125	1.293	0.468	4.452	6.420
PHMPC <sub>graph</sub>	(0.9, 10)	11.856	<b>5.009</b>	<b>2.861</b>	<b>2.036</b>	<b>5.052</b>	8.244
HMPC	(0.5, 10)	11.396	4.788	2.842	2.126	4.856	<b>9.460</b>
PHMPC <sub>concat</sub>	(0.5, 10)	10.533	3.809	1.616	0.961	4.509	6.678
PHMPC <sub>graph</sub>	(0.5, 10)	<b>12.473</b>	<b>5.566</b>	<b>3.510</b>	<b>2.725</b>	<b>5.120</b>	8.733

# Model Performance - Human

Models	Max	Human	HMPC	PHMPC <sub>concat</sub>	PHMPC <sub>graph</sub>
Relevance	1	0.766	0.266	0.133	<b>0.433</b>
Fluency	1	0.966	<b>0.566</b>	0.233	0.466
Informativeness	1	0.8	<b>0.166</b>	0.033	0.000
Utterance-level <sub>avg</sub>	3	2.533	<b>1.000</b>	0.400	0.900
Initiative-taking	1	0.700	<b>0.166</b>	0.000	0.100
Thread relevance	1	0.733	0.233	0.133	<b>0.366</b>
Persona relevance	1	0.733	0.366	0.266	<b>0.466</b>
Conversation-level <sub>avg</sub>	3	1.466	0.600	0.400	<b>0.833</b>

---

## Case Studies

- To observe whether the model performs well in conversations, we manually observe results on select conversation with differing properties such as political discussions
- Change the response speaker's persona for social media behavior to check if PersonaHeterMPC generations are relevant
- Eg, expressor → spectator

Speaker		Addressee		Utterance		
ID	Persona	ID	Persona	ID	Parent ID	Text
1	white male independent expressor	-1	-	0	-	when you can't take a joke.... <link>
2	white male liberal expressor	1	white male independent expressor	1	0	this is a mix of toxic masculinity and privilege (rich / famous) on display. the joke was in poor taste - yes. but resorting to violence to defend your wife from a joke. unacceptable. also, any other person (not rich / famous) would have been asked to leave / arrested.
3	white male liberal expressor	1	white male independent expressor	2	0	i can only hope that it was a staged event and not real.
4	white female conservative expressor	1	white male independent expressor	3	0	which is why ricky gervais will probably never host anything again.
5	white male liberal spectator	1	white male independent expressor	4	1	pathetic display by will smith ( <b>Human</b> )
5	white male liberal spectator	1	white male independent expressor	4	1	i don't think we have to see how he're in this is in a lot of the country. ( <b>HMPC</b> )
5	white male liberal spectator	1	white male independent expressor	4	1	. is a lot of them. ( <b>PHMPC<sub>concat</sub></b> )
5	white male liberal spectator	1	white male independent expressor	4	1	it's not a good one! ( <b>PHMPC<sub>graph</sub></b> )
5	white male liberal expressor	1	white male independent expressor	4	1	that's right.. ( <b>PHMPC<sub>concat</sub></b> )
5	white male liberal expressor	1	white male independent expressor	4	1	it's not a lot of the same thing to be so, but they are so it. ( <b>PHMPC<sub>graph</sub></b> )
5	white male liberal suppressor	1	white male independent expressor	4	1	that't....'s not just like a lot, i don's a lot of the same people who is. ( <b>PHMPC<sub>concat</sub></b> )
5	white male liberal suppressor	1	white male independent expressor	4	1	it's just a lot of the real. ( <b>PHMPC<sub>graph</sub></b> )
5	white male liberal avoider	1	white male independent expressor	4	1	is the time to do. ( <b>PHMPC<sub>concat</sub></b> )
5	white male liberal avoider	1	white male independent expressor	4	1	they are right! ( <b>PHMPC<sub>graph</sub></b> )

---

# Case Studies Takeaways

- Generations vary with the speaker persona, and reflect it well
- **PersonaHeterMPC<sub>graph</sub>** performs better than PersonaHeterMPC<sub>concat</sub> - modeling persona as nodes is a better strategy
- Generated response is in keeping with political and thus emotional charge of the conversation as well as the speaker persona

---

## Results

- Our work provides a baseline for modeling **persona-aware** response generation towards open domain MPC modeling using heterogeneous graph transformer network
- Human evaluations show that modeling personas into the conversation graph makes generations more relevant and engaging
- Case studies show that interaction types can be extended to active-additive (type 1) interactions

---

## Limitations

- Dataset size is limited - 50x smaller than Ubuntu IRC
- Participants tend to utilize either text or images/gifs to respond in conversations - need to account for multimodal input
- General issues with NLG evaluation apply for MPC evaluation as well - large-scale empirical evaluation required towards better generalization

---

## Future Directions

- Synthetically augment existing dataset with LLMs
    - Shown to perform well to boost dataset size and quality
  - Larger scale empirical evaluation
  - Work towards benchmark for MPC evaluation
-

---

**Thank you!**

**Questions?**

**Email: [kmahaja2@uncc.edu](mailto:kmahaja2@uncc.edu)**