

Reflecting on Experiences for Response Generation

Chenchen Ye

National University of Singapore
chenchenye.ccy@gmail.com

Suyu Liu

Singapore Management University
liusuyu4@gmail.com

Lizi Liao*

Singapore Management University
liaolizi.llz@gmail.com

Tat-Seng Chua

Sea-NExT Joint Lab, National University of Singapore
dcscts@nus.edu.sg

Multimodal Dialogue Systems



Hey, I am traveling with a group of four people, looking for a moderate price restaurant. Can you help me?



Hi. Do you have preferred location?



Better be in downtown core. I would like to try some jamon serrano.



What type of food would you like?



I think western food will be good.



I recommend the botanic. The food there is nice. Here are some pictures.




Dialogue History/ Context


Pure Textual / Pure Visual Response


Both Textual & Visual Response

Multimodal Dialogue Systems - Challenges

Challenges

 Hey, I am traveling with a group of four people, looking for a moderate price restaurant. Can you help me?

 *Hi. Do you have preferred location?*

 Better be in downtown core. I would like to try some jamon serrano.

 *What type of food would you like?* 

 *That sounds good.* 


- **Context-specific responses**


- Towards **user-specific** requests


- Emphasize the context-response mapping over the **whole training corpus**
- Tend to assign high probabilities to **safe but universal** responses (Li et al., 2016)

Multimodal Dialogue Systems - Challenges

Challenges


 Hey, I am traveling with a group of four people, looking for a moderate price restaurant. Can you help me?

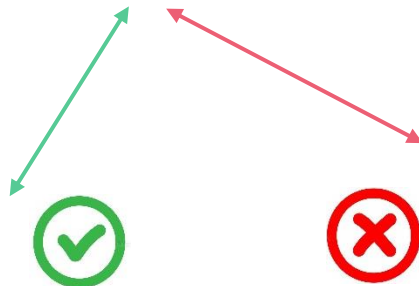
 *Hi. Do you have preferred location?*

 Better be in downtown core. I would like to try some jamon serrano.

 *What type of food would you like?*

 I think western food will be good.


 *I recommend the botanic. The food there is nice. Here are some pictures.*





- **Context-specific responses**
- **Coordination between the different modalities**
- **Coordinate** multimodal response components
- **Treat different modalities separately**

Multimodal Dialogue Systems - Challenges

Challenges


 Hey, I am traveling with a group of four people, looking for a moderate price restaurant. Can you help me?

 *Hi. Do you have preferred location?*

 Better be in downtown core. I would like to try some jamon serrano.

 *What type of food would you like?*

 I think western food will be good.

 *I recommend the botanic. The food there is nice. Here are some pictures.*



- **Context-specific responses**
- **Coordination between the different modalities**
- **Explainability**
- **Generalizability**

- **Why?**

- If we want to improve the responses for certain dialogue situation

-> Retraining & Catastrophic Forgetting

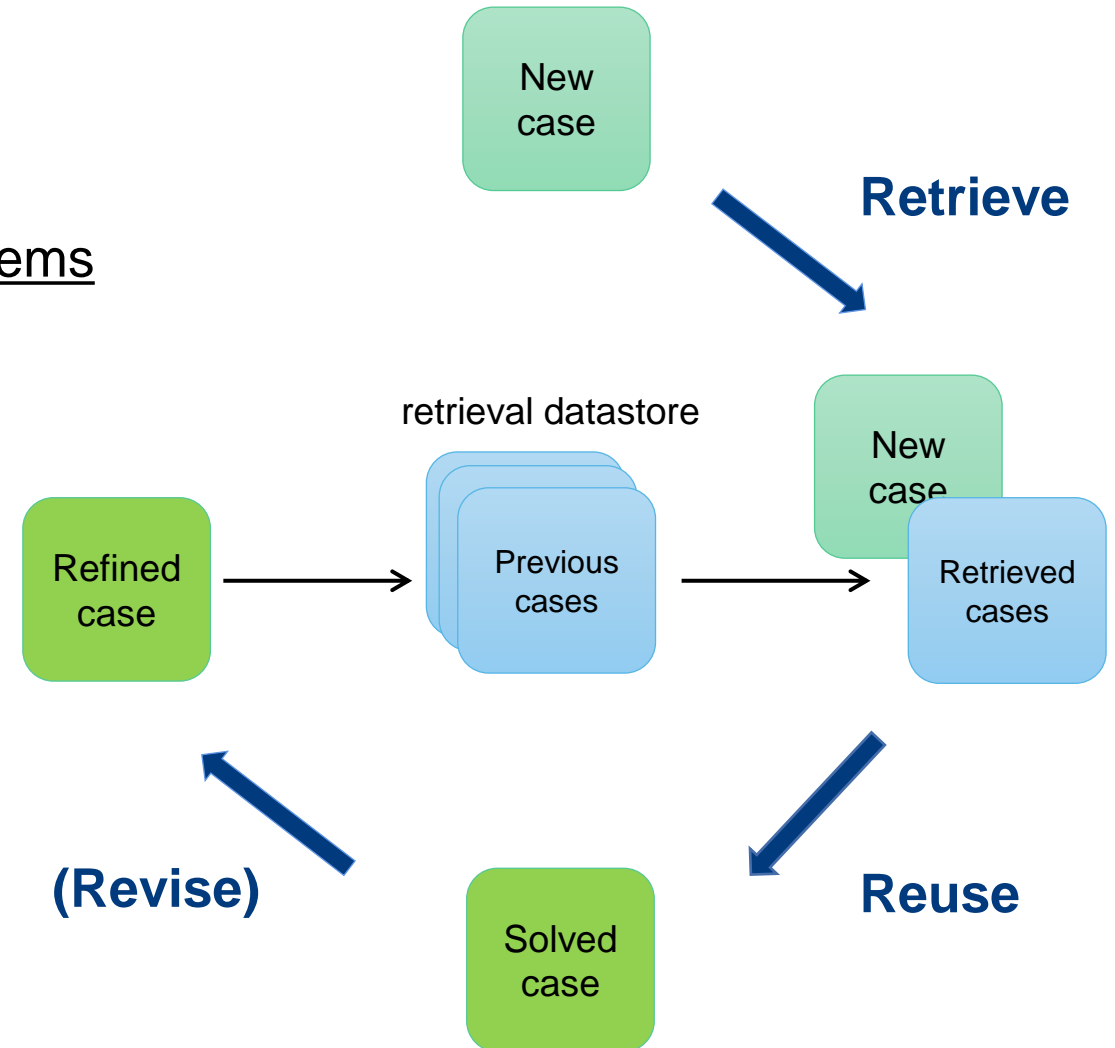
Case-based Reasoning (CBR)

Observation:

Humans often solve a new problem by **recollecting** and **adapting** the solutions to multiple related problems that they encountered in the past. (Ross, 1984)

A typical sketch of a CBR system:

- Retrieve
- Reuse
- Revise (if needed)



CBR – recent application in research

- **Knowledge-based question answering**
 - A Simple Approach to Case-Based Reasoning in Knowledge Base (Das et al., 2020)
 - CBR-KBQA (Das et al, 2021, EMNLP)
- **Natural language modeling**
 - kNN-LM (Khandelwal, 2020, ICLR)
- **Machine translation**
 - kNN-MT (Khandelwal, 2021, ICLR)
 - Adaptive kNN-MT (Zheng, 2021, ACL-IJCNLP)
- **Multimodal dialogue response generation**

structured triple queries /
pure textual sequences



complex dialogue queries /
multimodal cases

Similar Case 1

Similar Context

Hi! Can you help me look for a moderate price restuarant to try jamon serrano?

Anywhere in the region in particular?

Yes, the downtown core please.

Experienced Response

I would recommend the botanic! They have alot of great jamon serrano that you can try out along with other dishes.



Current Dialogue

Current Context

Hey, I am traveling with a group of four people, looking for a moderate price restaurant. Can you help me?

Hi. Do you have preferred location?

Better be in downtown core. I would like to try some jamon serrano.

What type of food would you like?

I think western food will be good.

Target Response

I recommend the botanic. The food there is nice. Here are some pictures.



Similar Case 2

Similar Context

Please reserve a place for me to try jamon serrano.

Sure, which region do you prefer?

Should be in downtown core. I am with a group of four people. Don't want to go far.

Experienced Response

Got it. The botanic is good. Great food with imaginative options, nice service, great ambience, good wine selection by glass and bottle. Try the anchovies on toast.



RERG

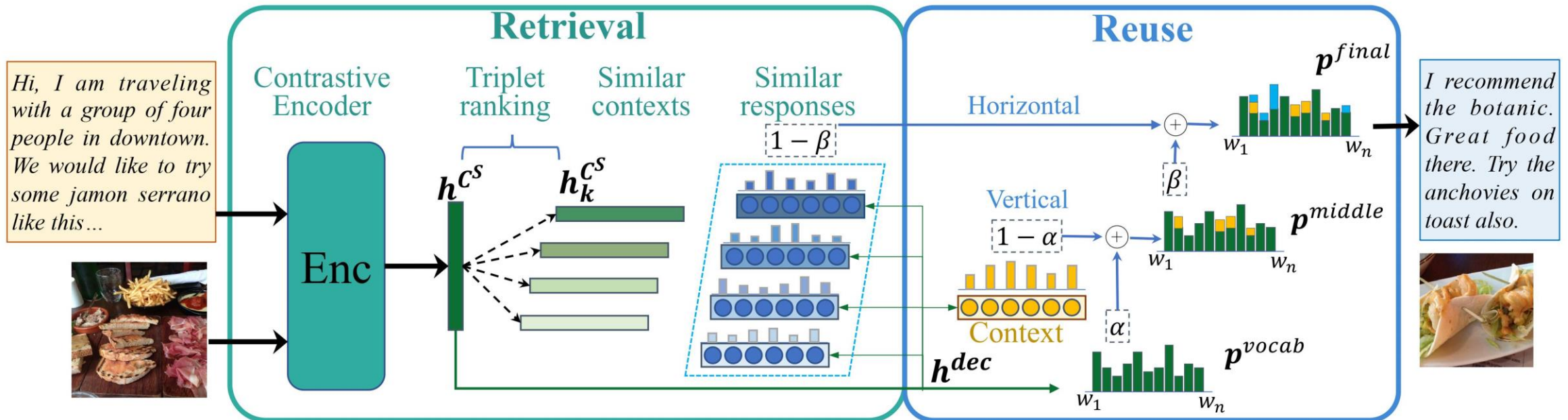
- A neural components based CBR framework to Reflect on Experiences for Response Generation

Retrieval Module

- Intra-modality Contrastive Learning
- Case-level Triplet Ranking

Reuse Module

- Reuse for Text Response – Cross Copy
- Reuse for Image Response



RERG – Retrieval Module – Intra-modality

- Textual Contrastive Learning for text context embedding s_i :

- Encode text context C_i^S with **BERT** as $s_i^{z_i} = f_s(C_i^S, z_i; \theta_s)$, where z^i is a random mask for dropout.
- Similar to SimCSE (Gao et al., 2021), the key here is to **get a positive embedding pair with different dropout masks** z^i and z^j .
- Training objective inside a minibatch of size N' is:

$$L_{\text{textual}} = -\log \frac{\exp(s_i^{z_i} \cdot s_i^{z'_i} / \tau)}{\sum_{j=0}^{N'} \exp(s_i^{z_i} \cdot s_j^{z'_j} / \tau)}$$

RERG – Retrieval Module – Intra-modality

- Visual Contrastive Learning for image context embedding q_i :
 - Augment the context image C_i^I to $C_i^{I'}$ and $C_i^{I''}$.
 - Following MoCo-v2 (Chen et al., 2020), embed them into **query** and **key** feature vectors by two **Resnet** network:

$$q_i = f_q(C_i^{I'}; \theta_q)$$
$$k_i = f_k(C_i^{I''}; \theta_k).$$

- Training objective with the **momentum queue** is:

$$L_{visual} = -\log \frac{\exp(q_i \cdot k_i^+ / \tau)}{\sum_{j=0}^M \exp(q_i \cdot k_i^j / \tau)}$$

RERG – Retrieval Module – Case-level

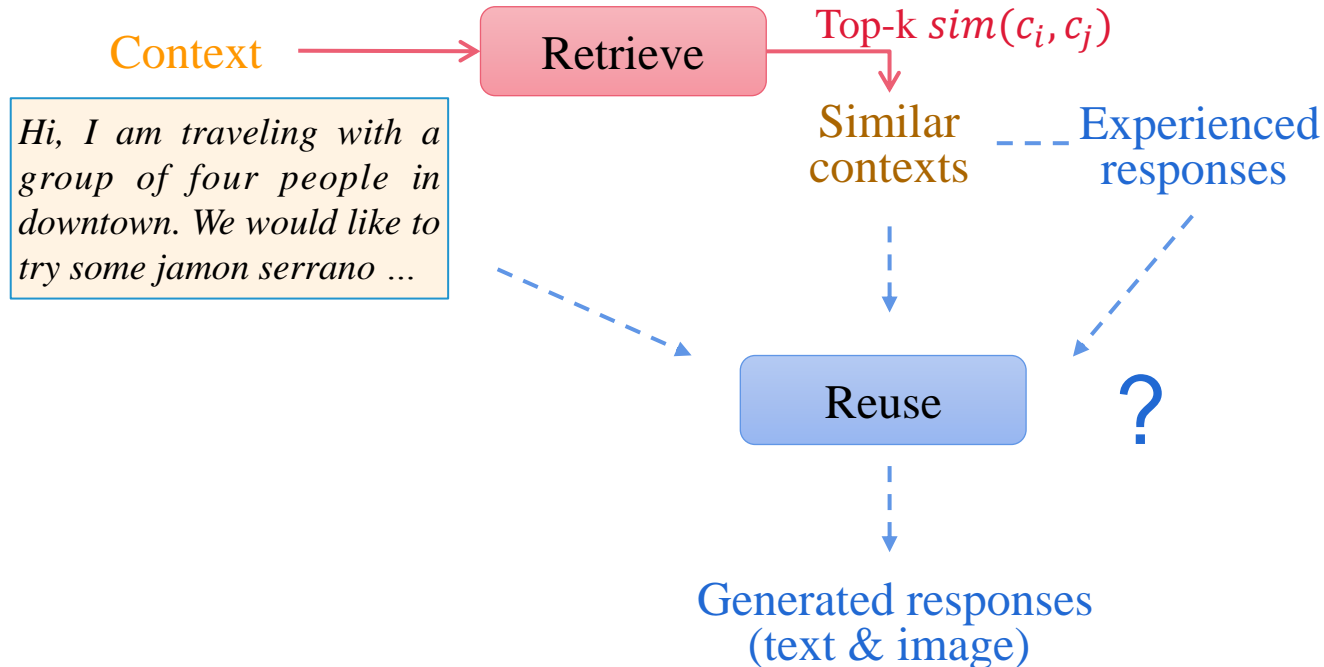
- Triplet Ranking for multimodal context representation c_i :
 - **Anchor**: Embed **current dialogue context** to c_i as $c_i = f_{MLP}([s_i; q_i])$
 - **Positive**: Encode **similar dialogue case's context** to c_i^+
 - **Negative**: Select the **batch-hardest case** embedding c_i^-
 - Triplet ranking loss with margin ϵ and **dot product** as the similarity function is:

$$L_{triplet} = \max(0, \epsilon - \text{sim}(c_i, c_i^+) + \text{sim}(c_i, c_i^-)).$$

RERG – Retrieval Module

- Intra-modality Contrastive Learning: Capture **intrinsic patterns** of text and image contexts
- Case-level Triplet Ranking: learn semantic information from **higher-level similarity**
- Thus, the **retrieval module** is trained via the **total loss** as:

$$L_{retrival} = L_{textual} + \lambda_1 \cdot L_{visual} + \lambda_2 \cdot L_{triplet}$$



RERG – Reuse Module – Text Response

Similar Case 1

Similar Context

Hi! Can you help me look for a moderate price restuarant to try jamon serrano?

Anywhere in the region in particular?

Yes, the downtown core please.

Experienced Response

I would recommend the botanic! They have alot of great jamon serrano that you can try out along with other dishes.



Current Dialogue

Current Context

Hey, I am traveling with a group of four people, looking for a moderate price restaurant. Can you help me?

Hi. Do you have preferred location?

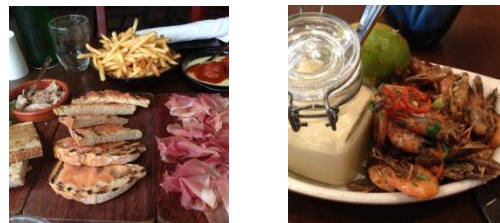
Better be in downtown core. I would like to try some jamon serrano.

What type of food would you like?

I think western food will be good.

Target Response

I recommend the botanic. The food there is nice. Here are some pictures.



Similar Case 2

Similar Context

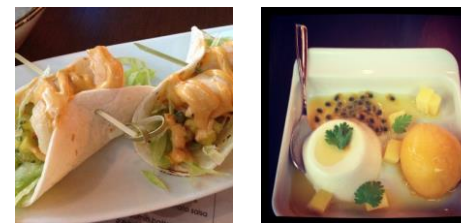
Please reserve a place for me to try jamon serrano.

Sure, which region do you prefer?

Should be in downtown core. I am with a group of four people. Don't want to go far.

Experienced Response

Got it. The botanic is good. Great food with imaginative options, nice service, great ambience, good wine selection by glass and bottle. Try the anchovies on toast.



Cross Copy

- Vallina Generation
- Vertical Copy
- Horizontal Copy

RERG – Reuse Module – Text Response

- Reuse for Text Response – Cross Copy

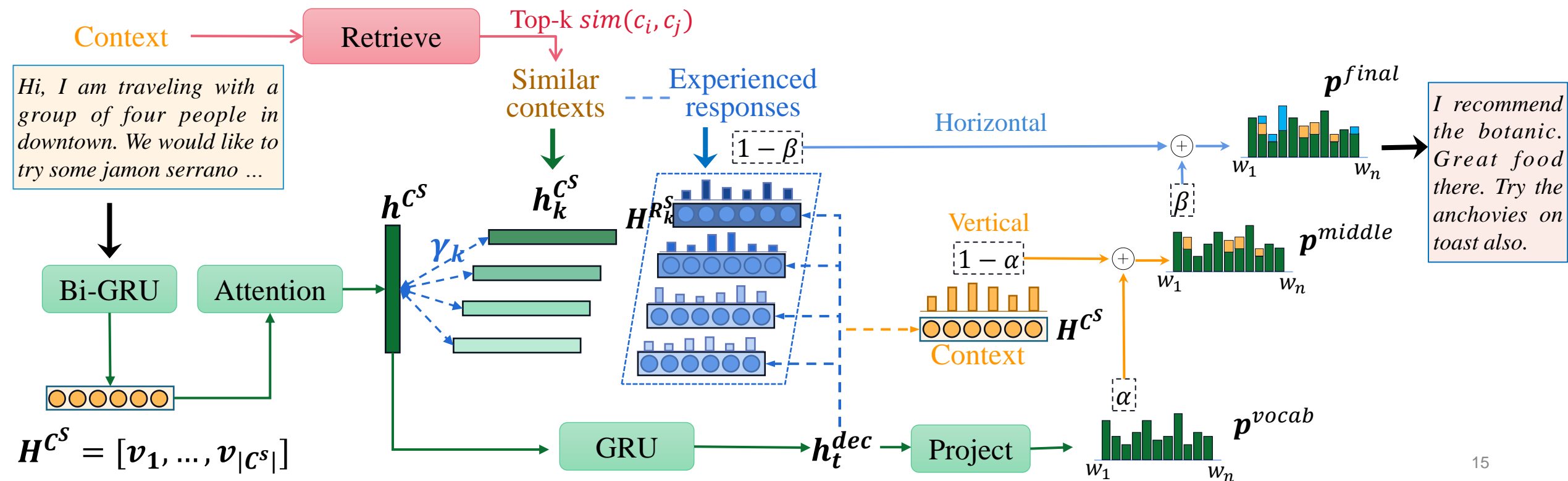
- **Vallina generation process:** a typical encoder-decoder network from context to response

- **Vertical Copy** from current dialogue context

$$\alpha = \text{sigmoid}(W_2 \cdot [h_t^{\text{dec}}; w_{t-1}; P_t^{\text{vertical}} \cdot H^{CS}])$$

- **Horizontal Copy** from the similar contexts' responses

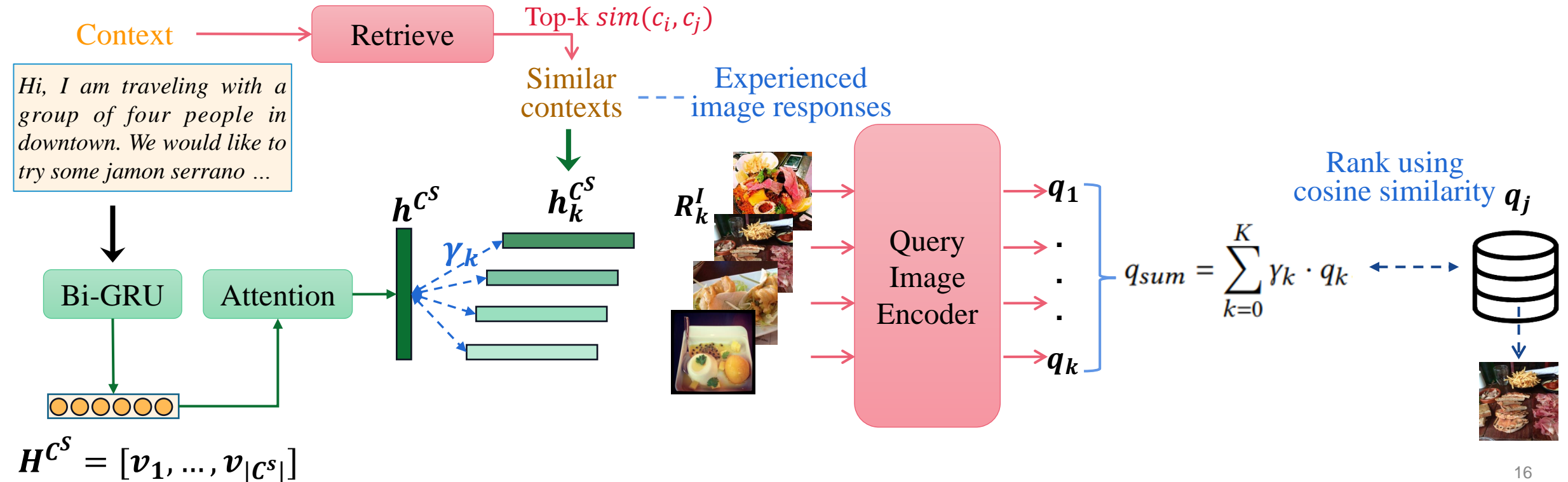
$$\beta_k = \text{sigmoid}(W_3 \cdot [h_t^{\text{dec}}; w_{t-1}; P_t^{\text{horizontal}} \cdot H^{RS_k}])$$



RERG – Reuse Module – Image Response

- Reuse for Image Response

- Weighted query vector q_{sum} by text context similarity γ_k
- Rank all images in datastore by the cosine similarity with q_{sum} , select the top-ranked image



RERG – Reuse Module

Similar Case 1

Similar Context

Hi! Can you help me look for a moderate price restuarant to try jamon serrano?

Anywhere in the region in particular?

Yes, the downtown core please.

Experienced Response

I would recommend the botanic! They have alot of great jamon serrano that you can try out along with other dishes.



Current Dialogue

Current Context

Hey, I am traveling with a group of four people, looking for a moderate price restaurant. Can you help me?

Hi. Do you have preferred location?

Better be in downtown core. I would like to try some jamon serrano.

What type of food would you like?

I think western food will be good.

Target Response

I recommend the botanic. The food there is nice. Here are some pictures.



Similar Case 2

Similar Context

Please reserve a place for me to try jamon serrano.

Sure, which region do you prefer?

Should be in downtown core. I am with a group of four people. Don't want to go far.

Experienced Response

Got it. The botanic is good. Great food with imaginative options, nice service, great ambience, good wine selection by glass and bottle. Try the anchovies on toast.



Cross Copy

- Vallina Generation
- Vertical Copy
- Horizontal Copy

Ranked Images

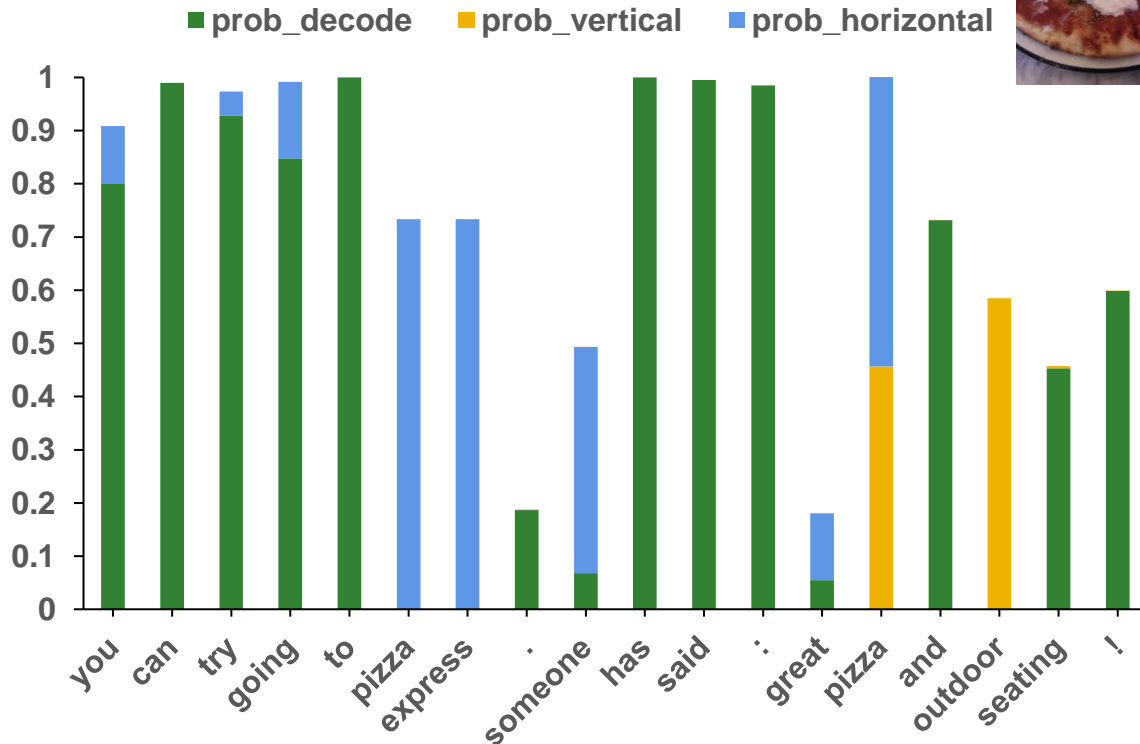
RERG – Explainability

Ground-truth
Response

You can try going to Pizza Express.

Generated
Response

*You can try going to Pizza Express.
Someone has said: Great pizza and outdoor seating!*



Context

Usr: I am thinking of a **pizza** place that has **outdoor** seating.

Sys: *Is there anything else you would like?*

Usr: I would like if they have moderate prices and if they accept credit cards.

Experienced Response

You can try going to Pizza Express. Someone has said: Great pizza, great service, tables to eat outside, perfect on a weekday evening after work. And I would recommend the pizza like in the picture.



*I would recommend their **pizza** like in the picture.*



Noted! In this case, I would recommend Pizza Express.

RERG – Performance

Dataset: MMConv (Liao et al, 2021), a multi-turn multimodal conversation dataset.

- 5,106 dialogues, 5 domains, 39.8K utterances
- Grounded on a venue database with 1,771 venues and 113,953 associated images

Main baseline model: employ the large pretrained GPT-2 model

Main multimodal response generation results

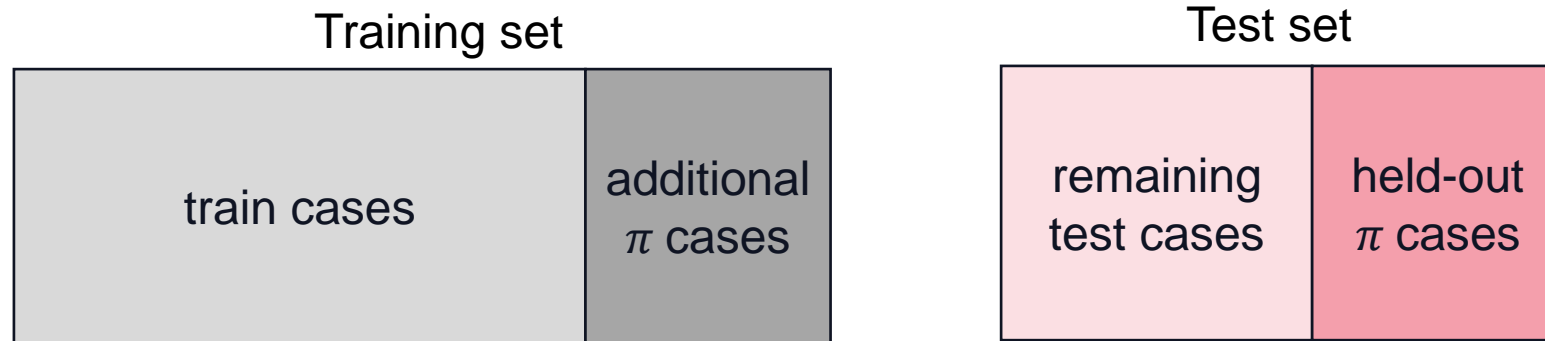
Group	Method	Textual Response					Image Response		
		BLEU	NIST	ROUGE-L	Entity-F1	Match Rate	Recall@1	Recall@3	Recall@5
Text-based	DialoGPT [46]	18.32	3.160	0.4419	18.89	24.7	–	–	–
	LaRL [50]	13.33	2.496	0.3214	5.36	1.5	–	–	–
	HDNO [40]	14.79	2.745	0.3663	8.23	2.3	–	–	–
Multimodal	MMD [37]	16.60	3.062	0.3728	11.08	5.1	4.69	8.33	11.98
	MMConv [23]	32.33	5.758	0.5402	49.01	69.2	17.85	–	–
	RERG_5	30.75	5.616	0.5585	52.55	79.3	22.83	24.88	26.33
	RERG _{gt_k=5}	31.17	5.529	0.5776	54.36	80.6	23.43	25.60	36.57
	RERG_2	29.66	5.374	0.5591	51.55	81.9	33.94	35.51	36.23
	RERG_10	27.72	5.345	0.5322	46.69	69.8	14.37	16.67	17.75

Natural pattern Targeted response Related images

RERG – Generalizability

Study on Unseen situations

- Consider all dialogues happened under the **user goal π** : “*You plan to do shopping in Jurong East. Thus seek about shopping malls there (Westgate, Jem, and IMM)*” .



Existing: **Time-consuming re-training** or **finetuning** process to handle unseen situations.

Such costly process may also lead to **catastrophic forgetting**.

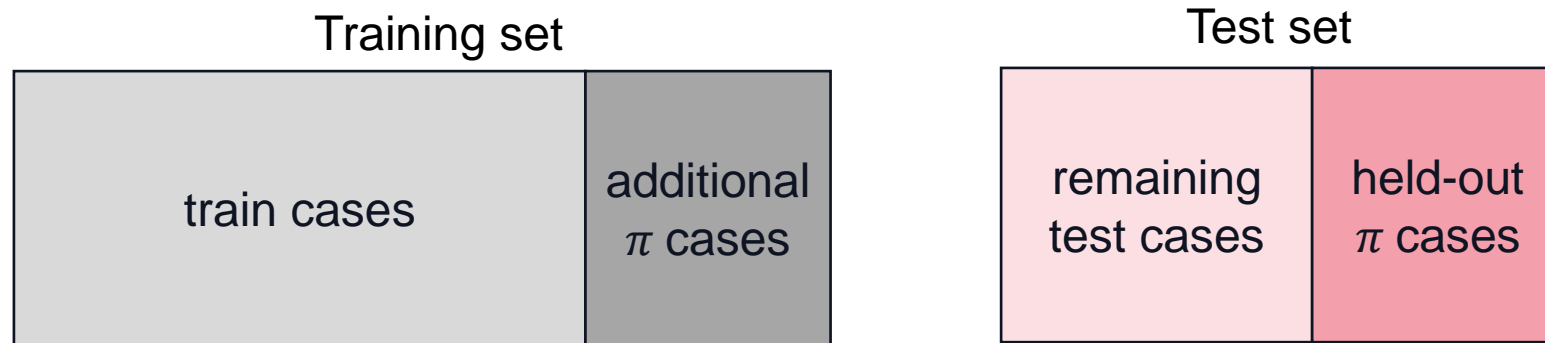
RERG: A **computationally much cheaper** way:

- add similar cases to the retrieve datastore
- let the reuse module to construct responses with new top-k cases.

RERG – Generalizability

Study on Unseen situations

- Consider all dialogues happened under the **user goal π** : “*You plan to do shopping in Jurong East. Thus seek about shopping malls there (Westgate, Jem, and IMM)*” .



Entity F1 score: task completion

Method	Scenario	Remaining	Held-out
MMConv	Train on original cases	49.07	11.54
	+ Fine-tune on additional cases	44.06	69.23
	+ Fine-tune on all cases	47.39	57.69
RERG	Train on original cases	49.55	11.54
	+ Add back to retrieve datastore	49.55	65.38

catastrophic forgetting

without retraining

RERG: A neural case-based reasoning framework to reflect on experiences for multimodal response generation

- Context-specific response to fulfill user requests
- Coordination between modalities
- Explainability
- Generalizability

Thank you for listening!

Email: chenchenye.ccy@gmail.com