对话系统的可控性

——对话系统中的个性化回复生成与异常点检测

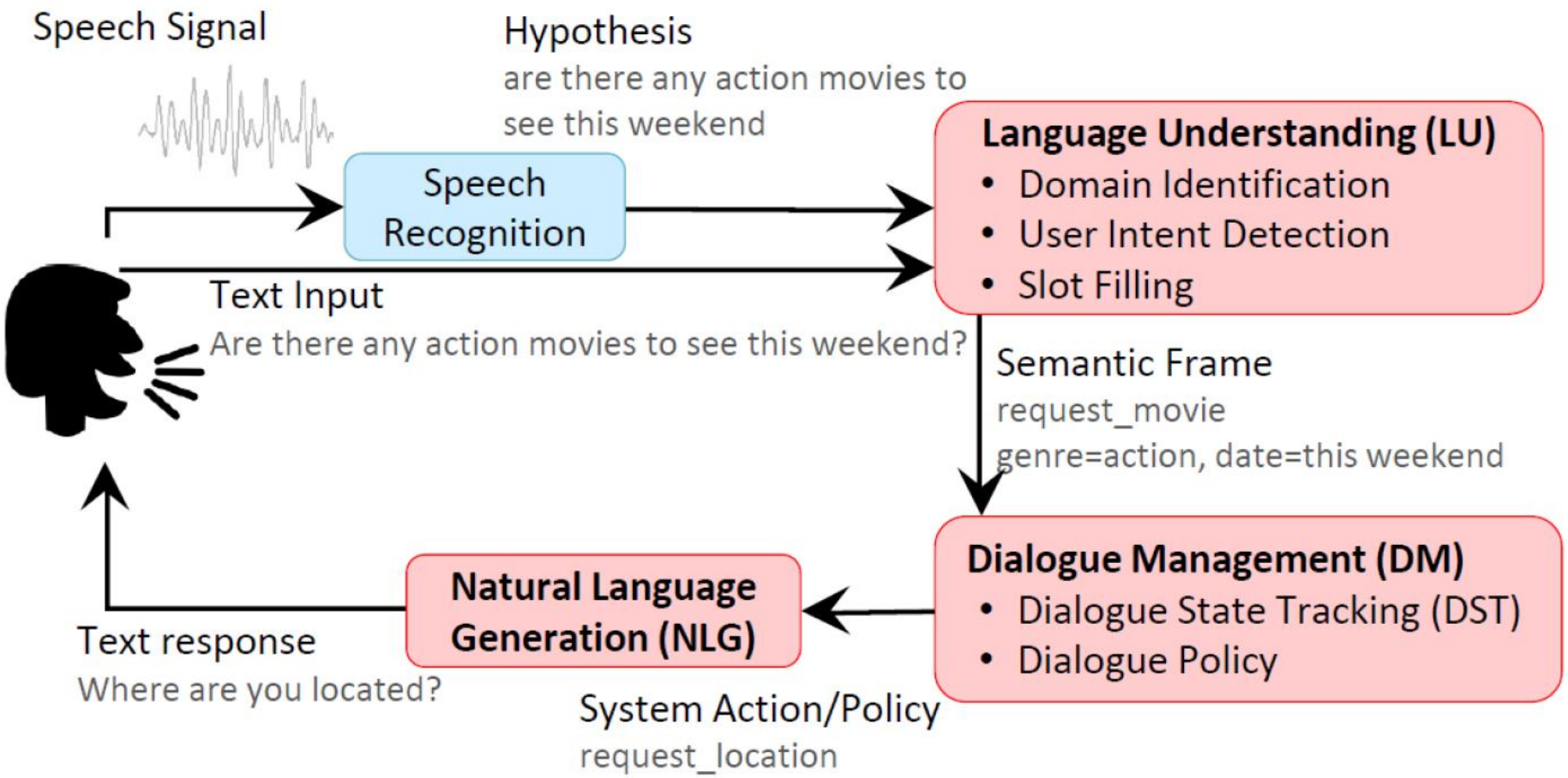
郑银河

清华大学交互式人工智能（CoAI）课题组

目录

- 对话系统中的NLU和NLG
- NLG中的个性化回复生成
- NLU中的异常输入检测(Out-of-Domain Detection)

对话系统中的NLU和NLG



个性化回复生成

Persona 1	Persona 2
I like to ski	I am an artist
My wife does not like me anymore	I have four children
I have went to Mexico 4 times this year	I recently got a cat
I hate Mexican food	I enjoy walking for exercise
I like to eat cheetos	I love watching Game of Thrones

[PERSON 1:] Hi
[PERSON 2:] Hello ! How are you today ?
[PERSON 1:] I am good thank you , how are you.
[PERSON 2:] Great, thanks ! My children and I were just about to watch Game of Thrones.
[PERSON 1:] Nice ! How old are your children?
[PERSON 2:] I have four that range in age from 10 to 21. You?
[PERSON 1:] I do not have children at the moment.
[PERSON 2:] That just means you get to keep all the popcorn for yourself.
[PERSON 1:] And Cheetos at the moment!
[PERSON 2:] Good choice. Do you watch Game of Thrones?
[PERSON 1:] No, I do not have much time for TV.
[PERSON 2:] I usually spend my time painting: but, I love the show.

Open-domain Dialogues are **Persona-sparse**

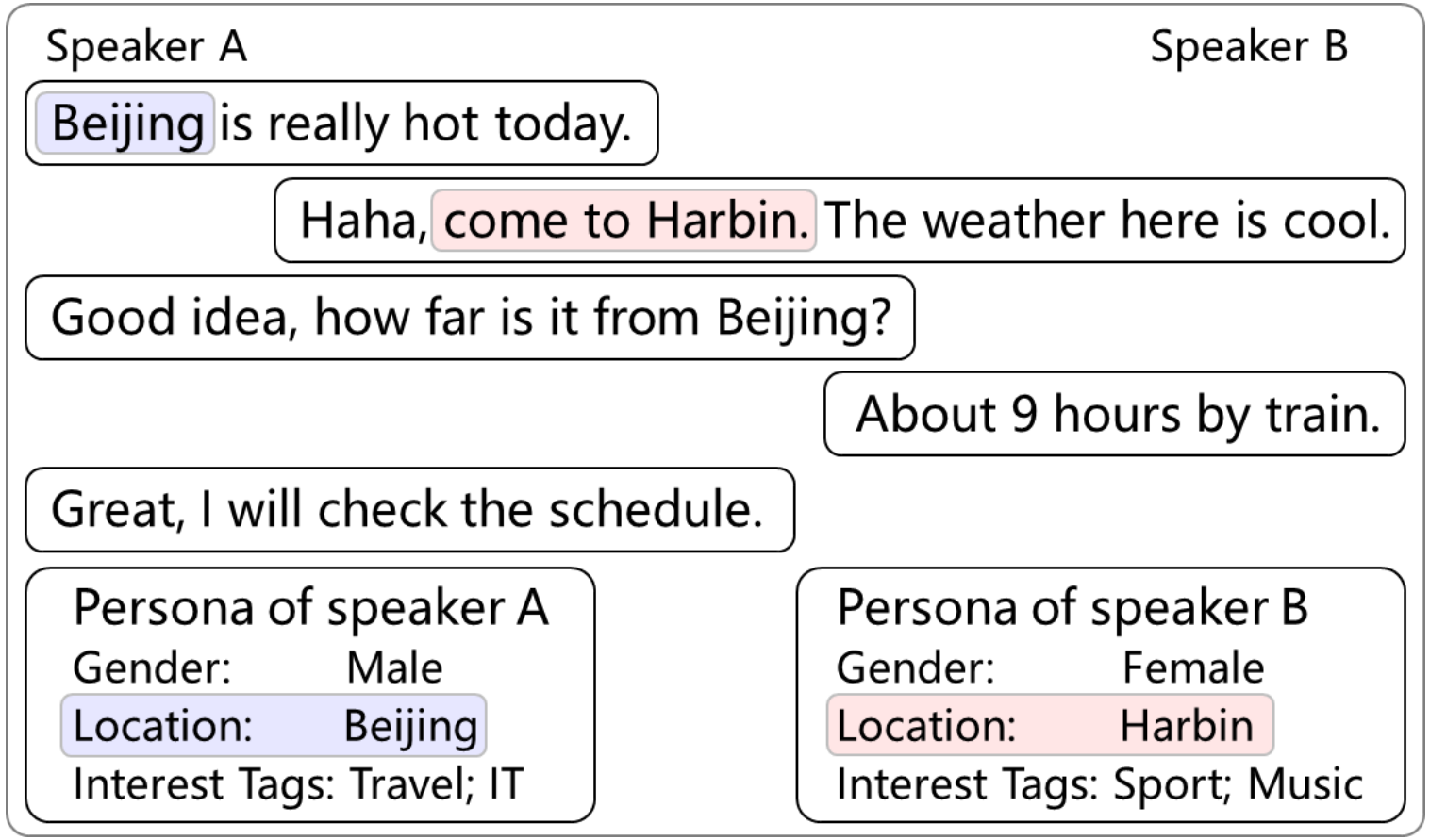
Persona related

Persona related

Non-persona related

Non-persona related

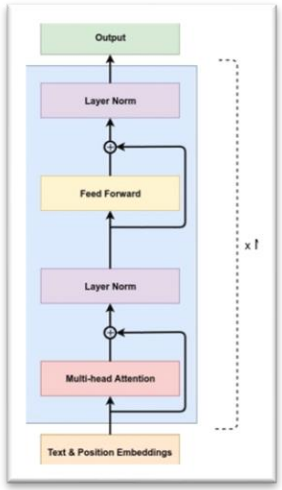
Non-persona related



An example dialogue sampled from the **PersonalDialog** dataset

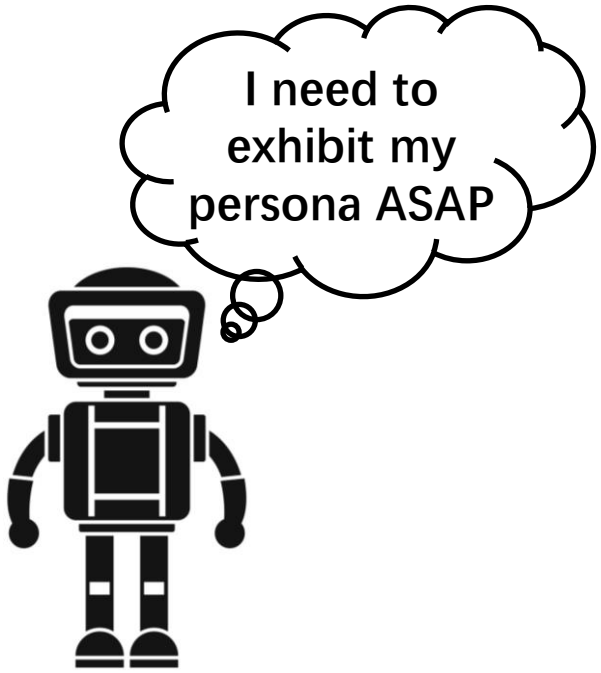
Problems When Using **Persona-dense** Data

How **PERSONA-CHAT** is collected



Pre-trained GPT

Fine-tune with **PERSONA-CHAT**

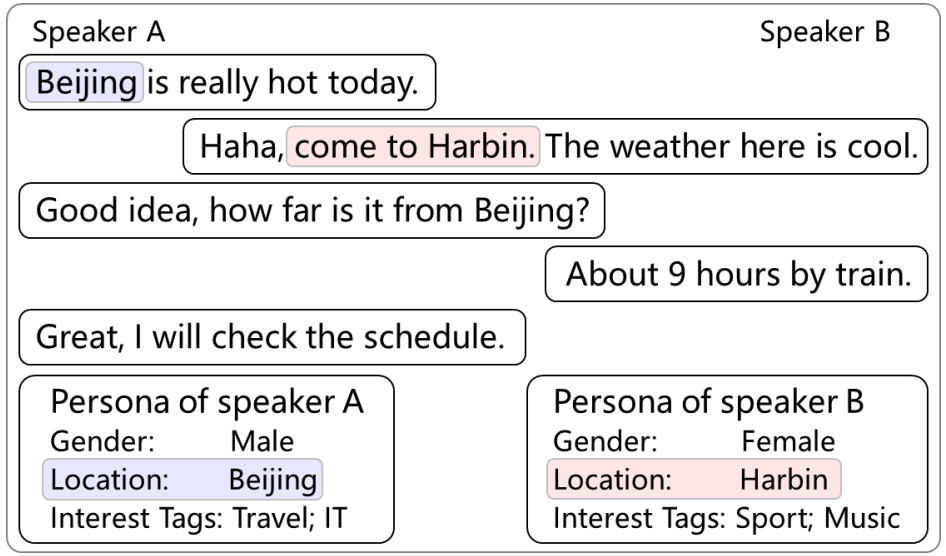


Direct fine-tuning lead to in-coherent responses

Any Solution?

- A pre-training-based dialogue model that can utilize persona-sparse dialogue data
- An attention routing mechanism to control the amount of persona-related features to exhibit in the response

Task Formulization



Inputs:

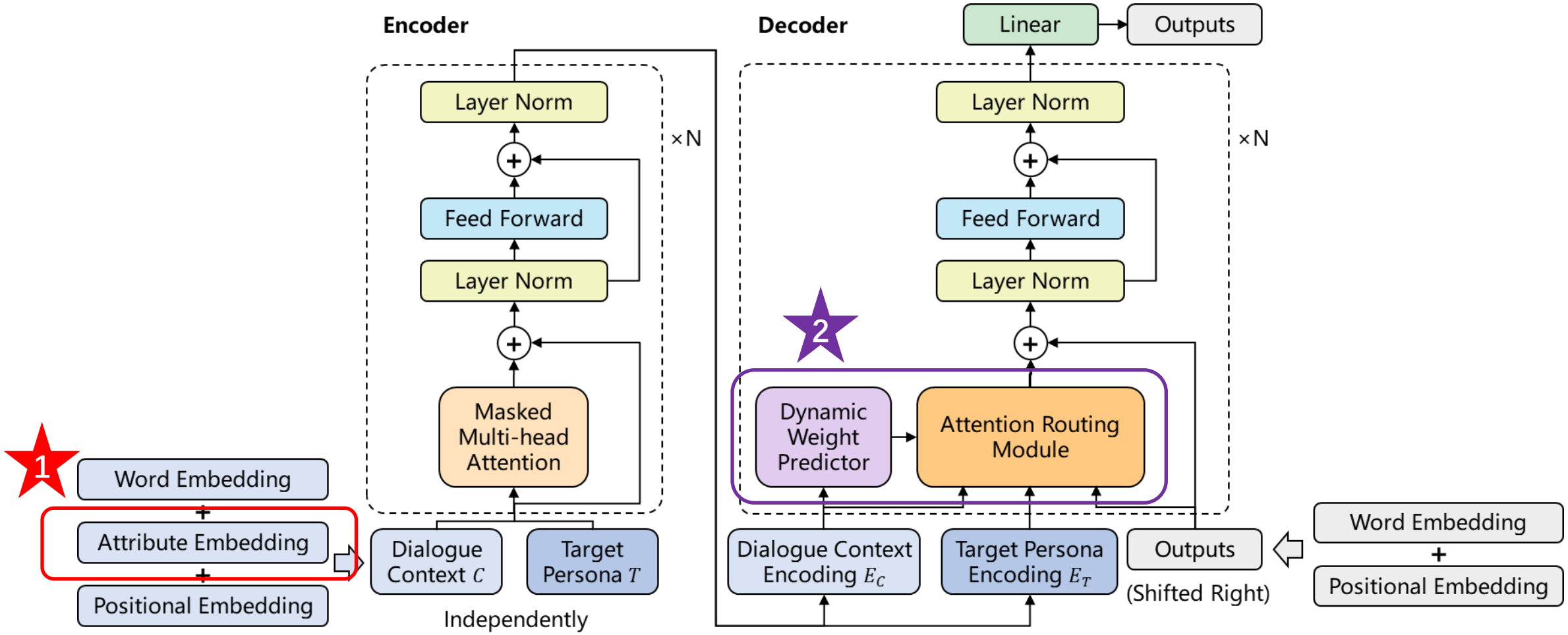
- Post X
- Responder's persona profile T: $\langle k_1, v_1 \rangle, \langle k_2, v_3 \rangle, \ldots, \langle k_N, v_N \rangle$

Output:

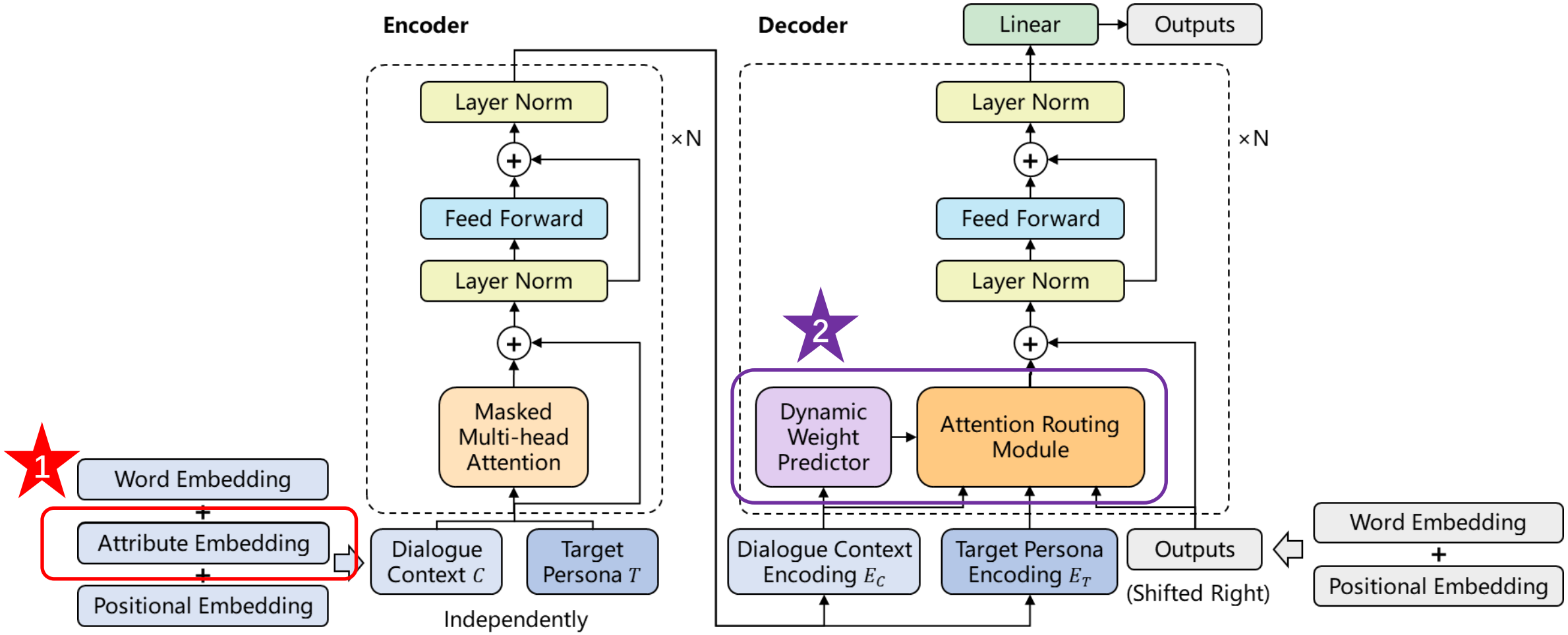
- Response Y, such that

$$Y = \arg\max_{Y^*} P(Y^* | X, T)$$

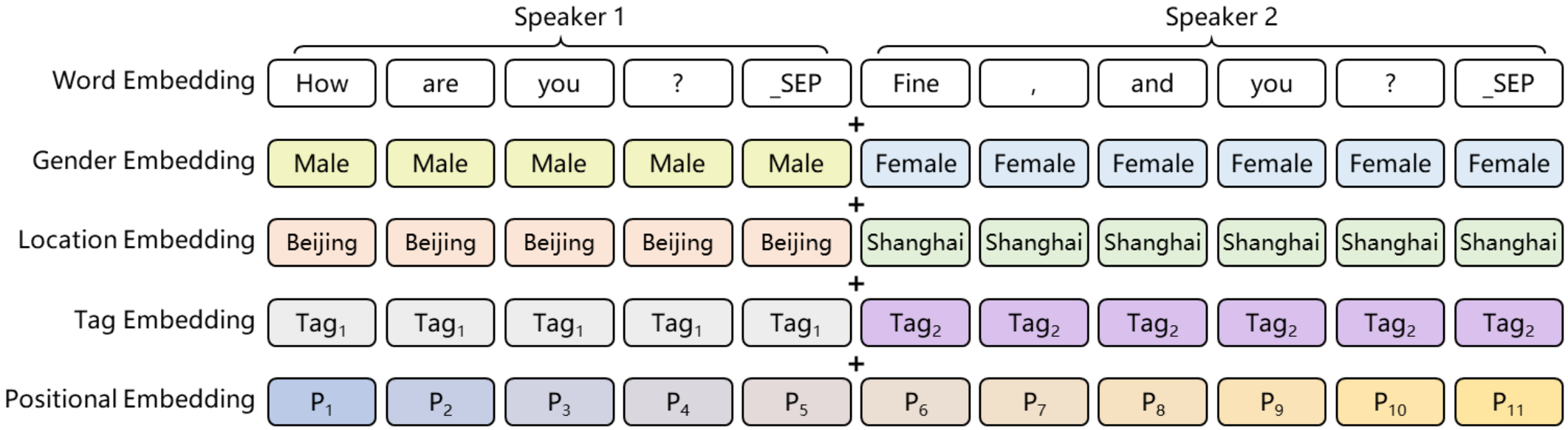
Model: Overview



Model: Overview



Model: Attribute Embedding

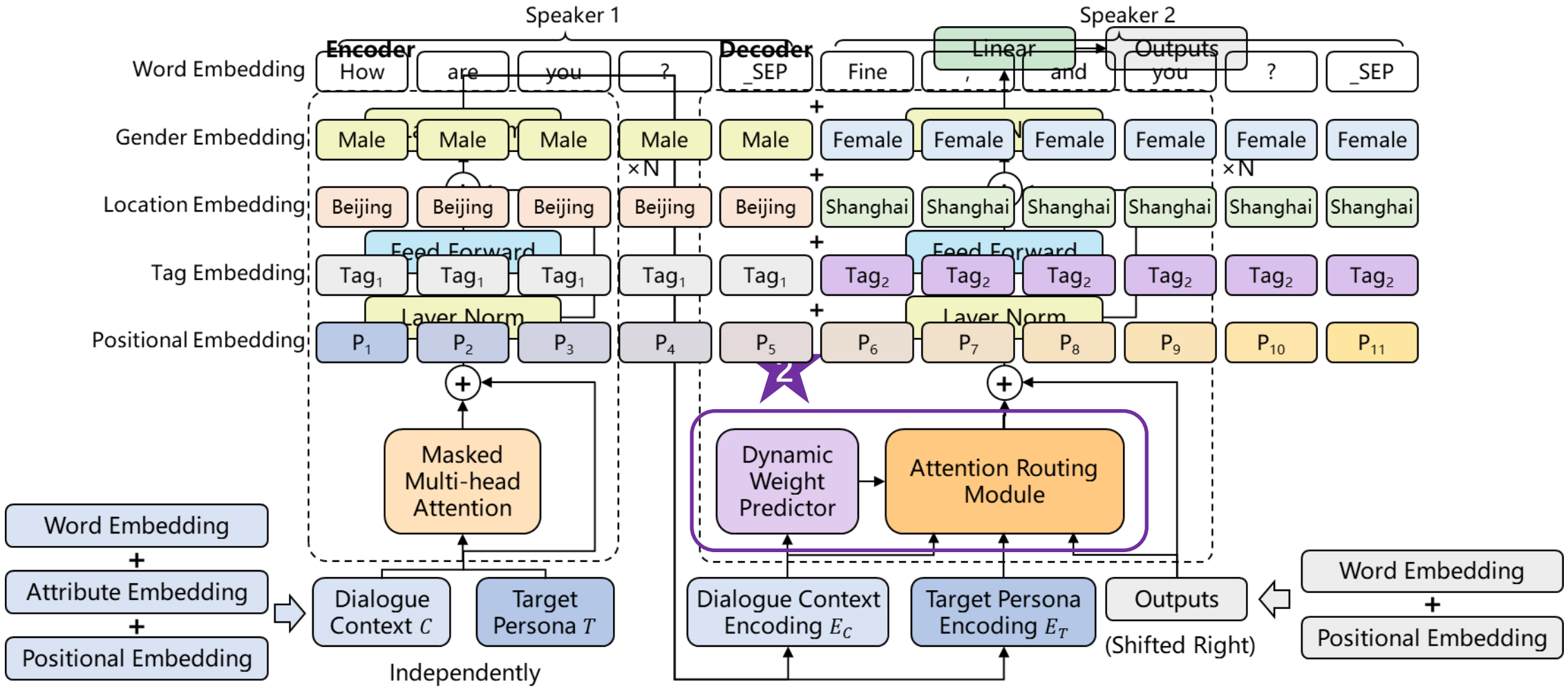


The input representation of the dialogue context

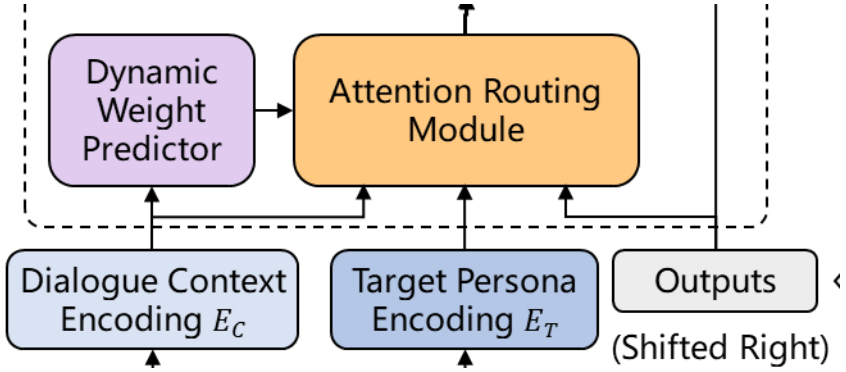
The input embedding for each token is the sum of

1. A word embedding
2. A positional embedding
3. Attribute embeddings (Gender embeddings, Location embeddings, Tag embeddings)

Model:


Speaker 1
Speaker 2
Encoder
Decoder
Linear
Outputs
Word Embedding
Gender Embedding
Location Embedding
Tag Embedding
Positional Embedding
How
are
you
?
_SEP
Fine
,
and
you
?
_SEP
Male
Female
Beijing
Shanghai
Tag1
Tag2
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
×N
Feed Forward
Layer Norm
Masked Multi-head Attention
Dynamic Weight Predictor
Attention Routing Module
Word Embedding
+
Attribute Embedding
+
Positional Embedding
Dialogue Context C
Target Persona T
Independently
Dialogue Context Encoding E_C
Target Persona Encoding E_T
Outputs
(Shifted Right)
Word Embedding
+
Positional Embedding

Model: Attention Routing



Model: Attention Routing

Dynamic Weight predictor

$$\alpha = P_\theta(\text{is persona related}|E_C)$$

Attention Routes

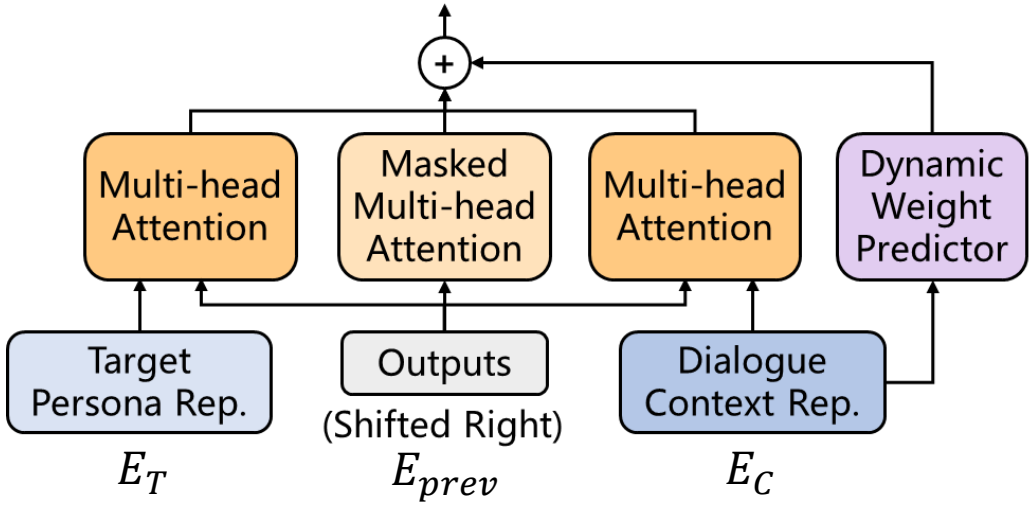
Query Key Value

$$O_T = MultiHead(E_{prev}, E_T, E_T)$$

$$O_c = MultiHead(E_{prev}, E_c, E_c)$$

$$O_{prev} = MultiHead(E_{prev}, E_{prev}, E_{prev})$$

$$O_{merge} = \alpha O_T + (1-\alpha)O_C + O_C + O_{prev}$$



The weight predictor and three attention routes

Model: Loss and Training

Loss:

Predictor loss: $$L_W(\theta) = -\sum_i r_i \log P_\theta(r_i|E_C) + (1 - r_i)\log[1 - P_\theta(r_i|E_C)]$$

Language model loss: $$L_{LM}(\phi) = -\sum_i \log P_\phi(u_i|u_{i-k}, \dots, u_{i-1})$$

Dialogue model loss: $$L_D(\phi) = -\sum_i \log P_\phi(u_i|u_{i-k}, \dots, u_{i-1}, E_C, E_T)$$

Total loss: $$L(\phi, \theta) = L_D(\phi) + \lambda_1 L_{LM}(\phi) + \lambda_2 L_W(\theta)$$

Training:

- The encoder and decoder share a same set of parameters
- Parameters are initialized using a pretrained GPT model

Dataset: PersonalDialog

1. Three persona attributes were approached (“Gender”, “Location”, and “Interest Tags”)

2. Two test sets are used: a random test set and a biased test set

Table 1: Statistics of the dialogue dataset used in this study.

Total number of dialogues	5.44 M
Total number of speakers	1.31 M
Total number of utterances	14.40 M
Dialogues with more than 4 utterances	0.81 M
Average utterances per dialogue	2.65
Average tokens per utterance	9.46

Automatic Evaluation

Metric:

(1) Persona Accuracy, (2) BLEU, (3) F1, (4) Distinct

Table 2: Automatic evaluation on the random test set.

Model	Acc.	BLEU	F1	Dist.	ppl.
Att+PAB	13.99	1.61	8.60	0.130	69.30
Trans.	7.80	3.97	12.51	0.132	43.12
TTransfo	8.80	4.06	12.63	0.169	**32.12**
TTransfo+P	43.05	3.44	11.28	0.158	43.78
LConv	9.45	4.19	12.99	0.157	32.64
LConv+P	48.00	3.56	11.46	0.136	42.00
Ours	32.80	4.18	12.52	**0.171**	35.06
Ours, α=1	**84.55**	3.45	10.96	0.154	38.56
Ours, α=0	12.90	**4.56**	**13.02**	**0.171**	33.71
w/o PreT	27.10	3.86	11.62	0.146	48.48
w/o AEmb	31.85	4.15	12.56	0.164	35.75
w/o DWP	30.70	4.15	12.34	0.169	34.10
+ HW	32.55	3.50	11.90	0.151	38.52

Table 3: Automatic evaluation on the biased test set.

Model	Acc.	BLEU	F1	Dist.	ppl.
Att. + PAB	47.60	3.08	12.50	0.133	94.38
Trans.	34.93	7.06	15.38	0.203	85.80
TTransfo	45.87	8.68	17.39	0.260	**34.83**
TTransfo+P	61.61	9.10	18.41	0.257	38.07
LConv	44.34	8.47	17.08	0.238	37.44
LConv+P	59.88	9.82	18.91	0.231	41.68
Ours	92.13	10.53	19.47	0.256	38.68
Ours, α=1	**94.24**	**11.63**	**20.51**	**0.262**	39.74
Ours, α=0	51.44	9.00	17.44	0.249	40.89
w/o PreT	71.74	9.36	18.29	0.222	95.00
w/o AEmb	73.51	10.51	19.41	0.247	39.36
w/o DWP	73.90	10.61	19.26	0.256	37.08
+ HW	69.87	9.01	19.81	0.232	36.37

Manual Evaluation

Metric:

(1) Utterance Fluency, (2) Persona Consistency, (3) Context Coherency

Table 4: Manual evaluation on the random and biased test sets.

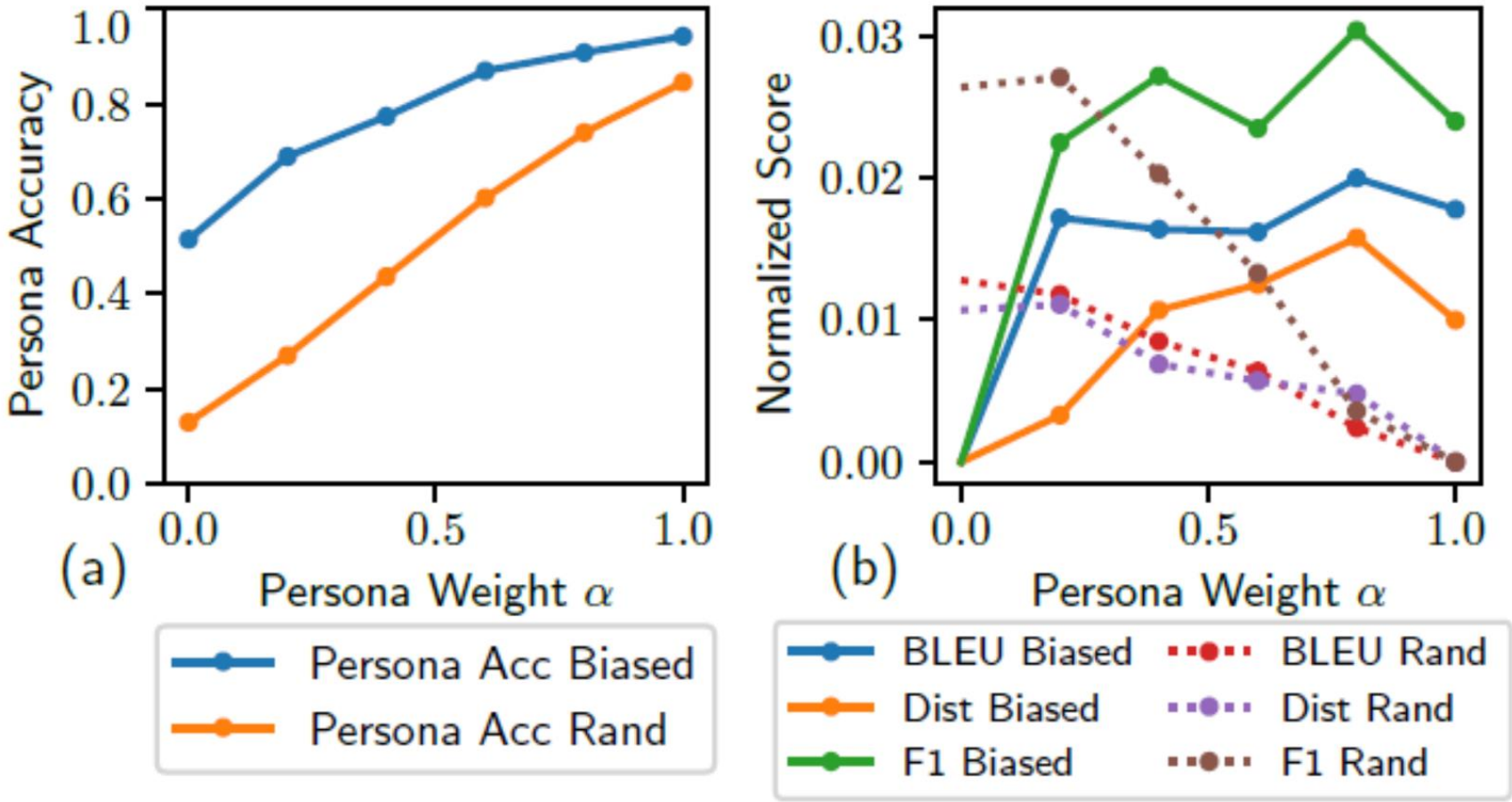
Model	Utterance Fluency		Persona Consistency		Context Coherency	
	Rand	Biased	Rand	Biased	Rand	Biased
Trans.	1.852	1.810†	0.997†	1.068†	1.428†	1.500
TTransfo	1.832†	1.890	1.015†	1.100†	1.498	1.517
TTransfo+P	1.802†	1.837†	1.125†	1.195†	1.217†	1.483†
LConv	1.863	1.882	1.028†	1.147†	1.490	1.550
LConv+P	1.832†	1.875†	1.093†	1.173†	1.238†	1.478†
Ours	1.837†	**1.912**	1.092†	1.198†	1.487	**1.563**
Ours, α=1	1.835†	1.900	**1.248**	**1.268**	1.303†	1.467†
Ours, α=0	**1.890**	1.880†	0.997†	1.085†	**1.535**	1.463†
Gold Resp	1.928	1.922	1.015	1.423	1.758	1.807

† significant difference with the best result (t-test, p-value<0.05)

Dialogue Context:	I am free this weekend. It that OK? (我这周末闲，好不) (Gender: Female, Location: Nanjing, Interest Tags: Cate)
Target Persona:	(Gender: Male, Location: **Nanjing**, Interest Tags: Null)
Trans. :	I'm also free, but I don't know what's going on. (我也闲，但是我不知道怎么回事)
TTransfo :	Great, come here and we can have a meal together. (好呀，你来了我们一起吃个饭呗)
TTransfo+P :	I am free this weekend. It that Okay? (我这周末闲，好不好？)
LConv :	It is not OK. (不好)
LConv+P :	Are you in **Nanjing** or **Nanjing**? (你在**南京**还是**南京**啊？)
Ours :	Ok, are you coming to **Nanjing**? (好啊，你来**南京**吗？)
Ours, $\alpha = 1$:	Come to **Nanjing**, I will treat you to dinner and film. (来**南京**，我请你吃饭看电影)
Ours, $\alpha = 0$:	Ok, where are you? I will go back tomorrow. (好啊，你在那呢？我明天就回去了)
Gold Resp :	OK, come to **Nanjing**! (好，来**南京**！)

Figure 5: Sample responses generated by baselines and our model.

Effect of α



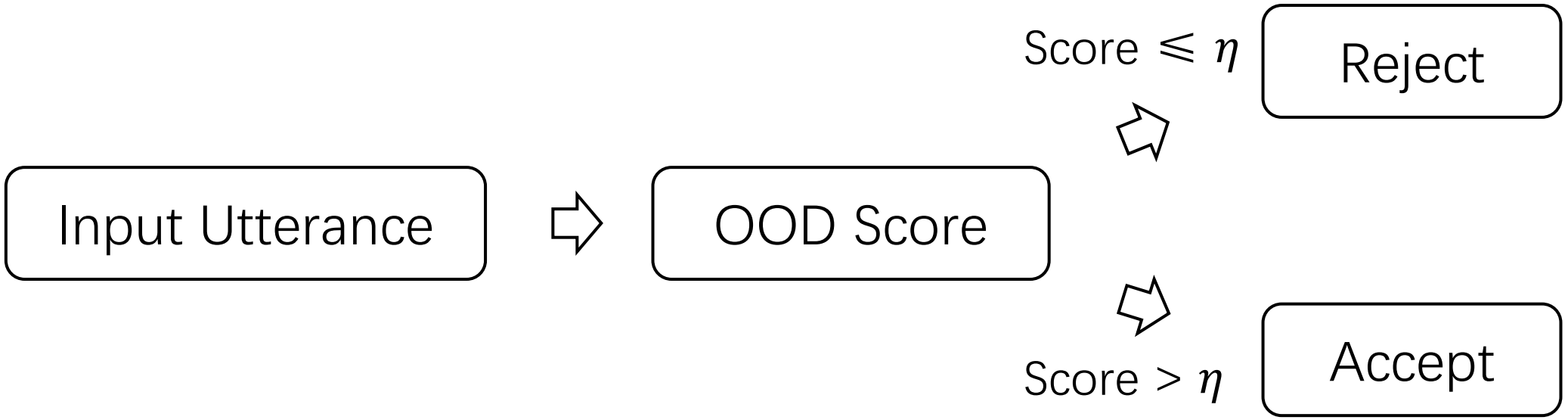
Conclusion

- Our model can produce coherent and persona-consistent responses
- The dynamic routing mechanism helps to utilize persona-sparse dialogues
- There are trade-offs between persona consistency and context coherency

NLU中的异常输入检测(Out-of-Domain Detection)

“把相册吃掉”

Threshold-based method



Threshold-based method

Basic solution: Use the maximum value of the Softmax outputs

$$Score(x) = \max_{i \in \{1,2,\ldots,m\}} P_\theta(y = l_i|x)$$

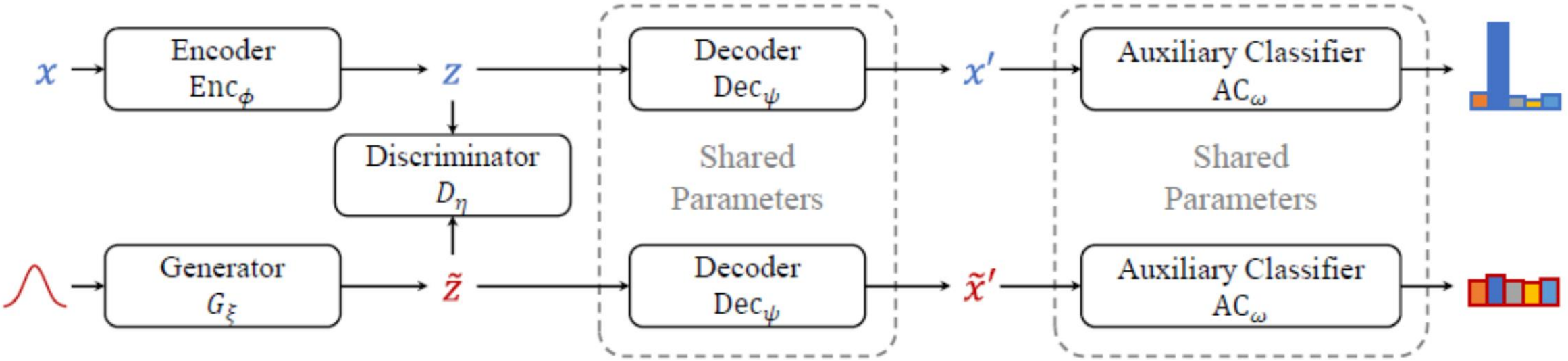
Advanced solution: Use an entropy regularization term

$$\mathcal{L}_{cls}(\theta) = \mathcal{L}_{ce}(\theta) + \alpha \mathcal{L}_{ent}(\theta)$$

$$\mathcal{L}_{ce}(\theta) = \mathop{\mathbb{E}}_{(x_i, y_i) \sim \mathcal{P}_{ind}} [-logP_\theta(y = y_i|x_i)]$$

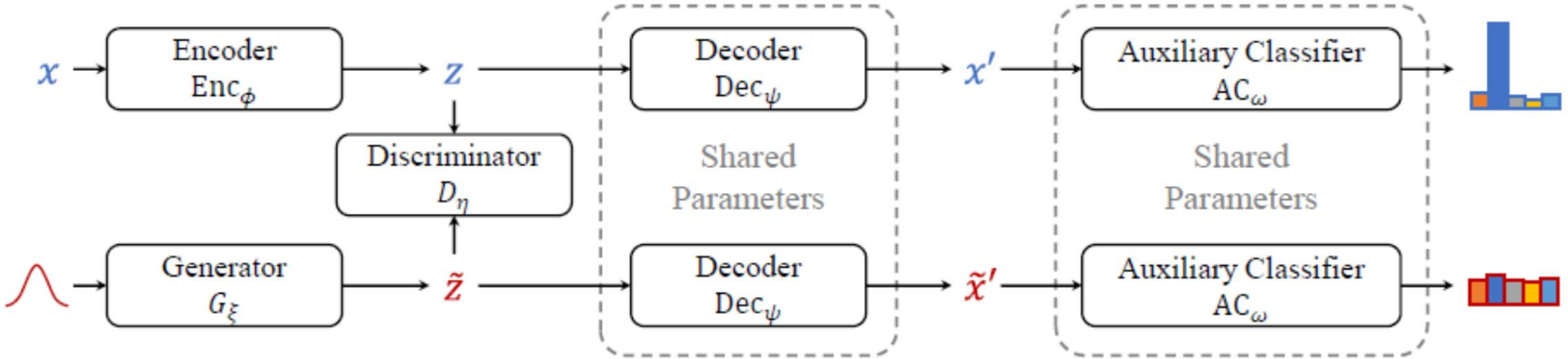
$$\mathcal{L}_{ent}(\theta) = \mathop{\mathbb{E}}_{\hat{x} \sim \mathcal{P}_{ood}} [-\mathcal{H}(P_\theta(y|\hat{x}))]$$

How to obtain OOD samples?



1. Autoencoder
2. Adversarial Generation Module
3. Auxiliary Classifier

How to obtain OOD samples?



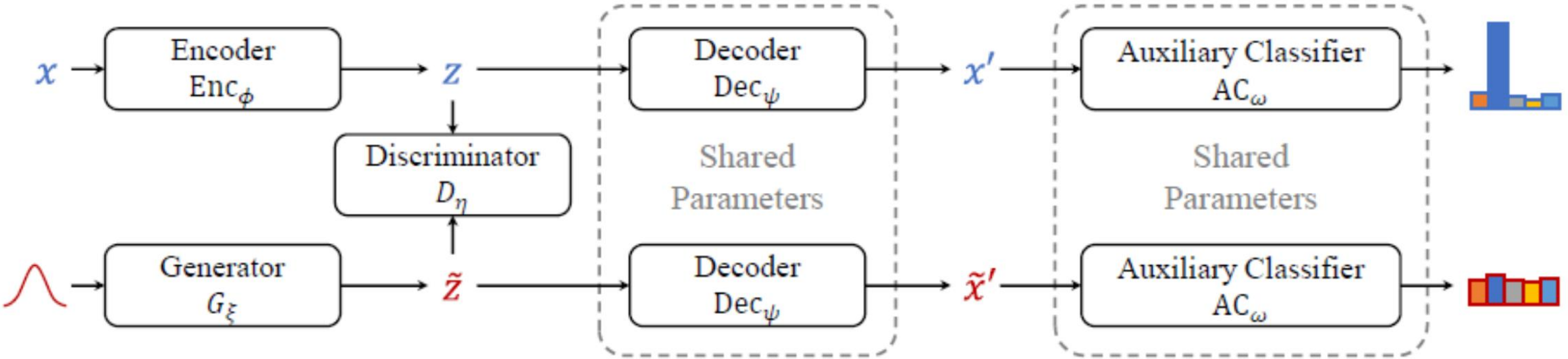
1. Autoencoder

$$z = \text{Enc}_\phi(x)$$

$$\mathcal{L}_{rec}(\phi, \psi) = \mathop{\mathbb{E}}_{\epsilon \sim \mathcal{N}(0,\mathbf{I})} \mathop{\mathbb{E}}_{x \sim \mathcal{P}_{ind}} [-\log P_\psi (x|z+\epsilon)]$$

$$= \mathop{\mathbb{E}}_{\epsilon \sim \mathcal{N}(0,\mathbf{I})} \mathop{\mathbb{E}}_{x \sim \mathcal{P}_{ind}} [-\log P_\psi \left(x\middle|\text{Enc}_\phi(x)+\epsilon\right)]$$

How to obtain OOD samples?



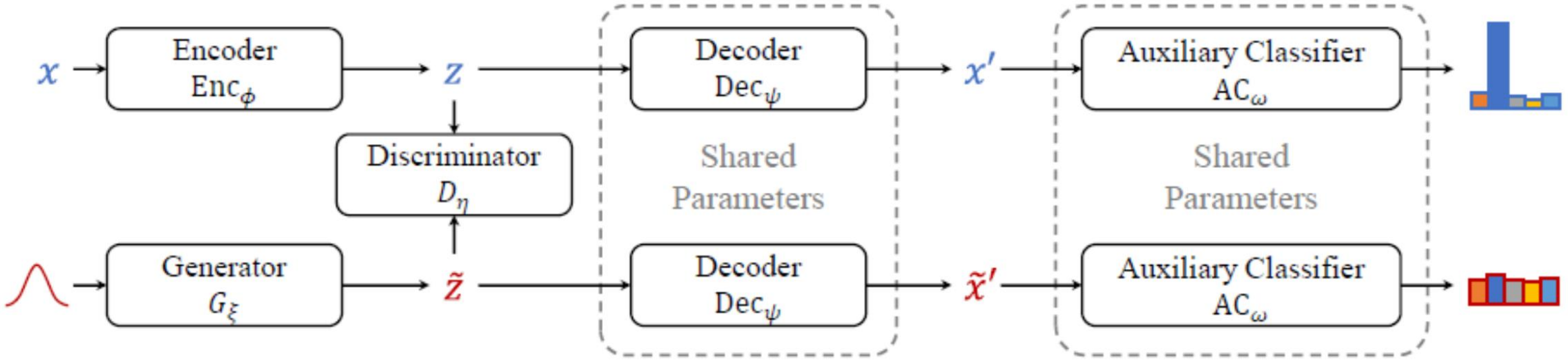
2. Adversarial Generation Module

$$\tilde{z} = G_\xi(\epsilon)$$

$$\mathcal{L}_g(\xi) = \mathop{\mathbb{E}}_{\epsilon\sim\mathcal{N}}[-D_\eta(G_\xi(\epsilon)]$$

$$\mathcal{L}_d(\eta) = \mathop{\mathbb{E}}_{\epsilon\sim\mathcal{N}}[D_\eta(G_\xi(\epsilon)] - \mathop{\mathbb{E}}_{x\sim\mathcal{P}_{ind}}[D_\eta(\text{Enc}_\phi(x))]$$

How to obtain OOD samples?



3. Auxiliary Classifier

$$\mathcal{L}'_{ce}(\omega) = \mathop{\mathbb{E}}_{\substack{(x_i,y_i)\sim\mathcal{P}_{ind}, \\ x'_i\sim P_\psi(x|\text{Enc}_\phi(x_i))}} [-logP_\omega(y=y_i|x'_i)]$$

$$\mathcal{L}'_{ent}(\xi) = \mathop{\mathbb{E}}_{\substack{\epsilon\sim N, \\ \tilde{x}'\sim P_\psi(x|G_\xi(\epsilon))}} [-\mathcal{H}(P_\omega(y|\tilde{x}'))]$$

Experiments

Datasets

		Train	Validate	Test
OSQ Dataset	IND	15.00K	3.00K	4.50K
	OOD	-	0.10K	1.00K
	$\mathcal{D}_{mix}$	10.25K	-	-
IPA Dataset	IND	28.90K	3.60K	3.60K
	OOD	-	1.20K	1.20K
	$\mathcal{D}_{mix}$	20.00K	-	-

Results on OSQ dataset

Model	AUROC↑	AUPR↑	FPR95↓	FPR90↓
Cont. GAN	52.22†	82.79†	94.40†	88.17†
Maha. Dis.	67.14†	90.56†	91.94†	83.30†
Likelihood Ratio	85.60†	96.07†	62.50†	43.40†
AE	87.78†	96.98†	58.50†	40.10†
MSP	92.86†	98.24†	39.72†	21.76†
Entropy	92.82†	98.87†	31.91†	19.64†
KNN	93.33†	98.20†	33.86†	18.78†
DOC	94.24†	98.55†	30.02†	14.94†
ODIN	95.14*	98.84*	26.04*	11.70†
ER+Perturb	94.01†	98.55†	34.04†	15.32†
ER+Mix	93.48†	98.31†	33.16†	17.86†
ER+POG	95.41*	98.94*	25.00*	10.10†
ER+AEPOG	**95.83**	**99.05**	**23.70**	**9.50**
w.o. Noise	94.03*	98.53*	35.22†	17.74†
w.o. Soft Token	93.85†	98.53*	39.24†	17.88†

Experiments

Results on IPA dataset (harder setting)

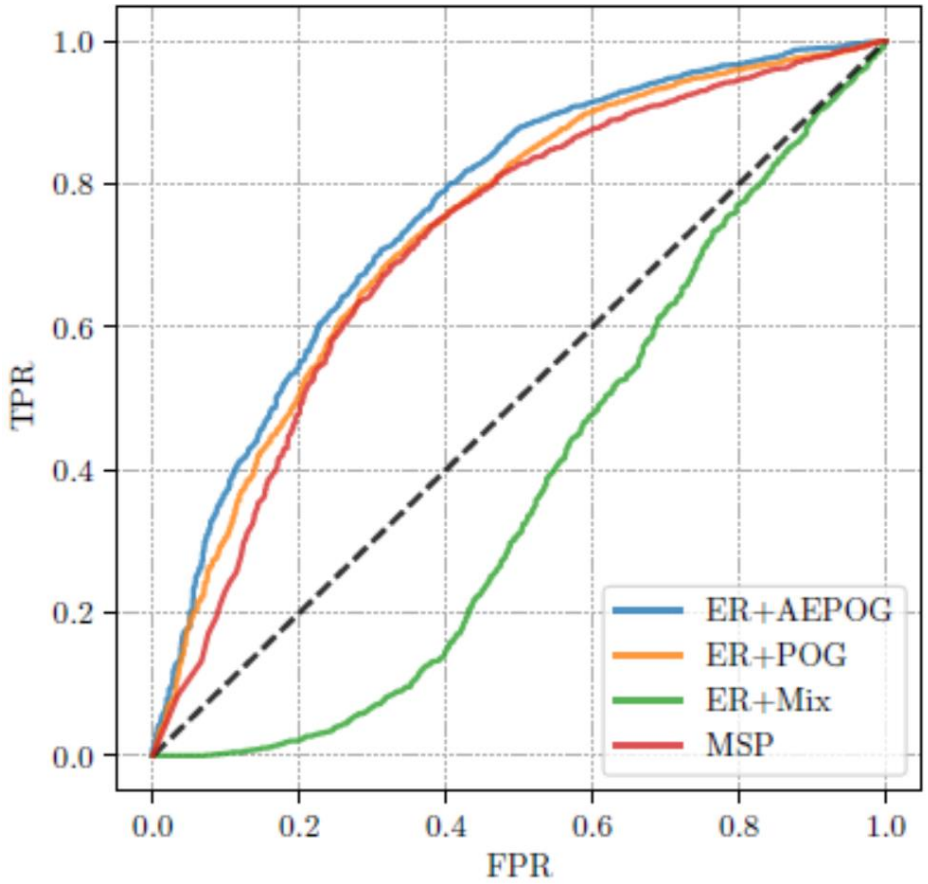
Model	AUROC↑	AUPR↑	FPR95↓	FPR90↓
Cont. GAN	58.06†	80.47†	91.51†	85.05†
Maha. Dis.	45.65†	23.45†	95.01†	89.97†
Likelihood Ratio	69.65†	86.21†	84.44†	74.46†
AE	67.57†	85.91†	86.77†	74.79†
MSP	72.66†	86.42†	77.79†	86.42†
Entropy	72.87†	86.42†	88.15†	75.41†
KNN	67.94†	82.62†	76.87†	63.43†
DOC	71.68†	46.03†	79.40†	64.34†
ODIN	72.90†	86.53†	77.57†	63.28†
ER+Perturb	73.24†	86.89*	77.88†	63.30†
ER+Mix	33.71†	63.55†	96.44†	92.33†
ER+POG	73.83†	87.15*	76.17†	62.28†
ER+AEPOG	**75.86**	**87.95**	**71.67**	**56.78**
w/o Noise	71.01*	85.10*	79.17†	65.17†
w/o Soft Token	72.09*	86.31*	80.22†	65.07†

Results on IPA dataset (easier setting)

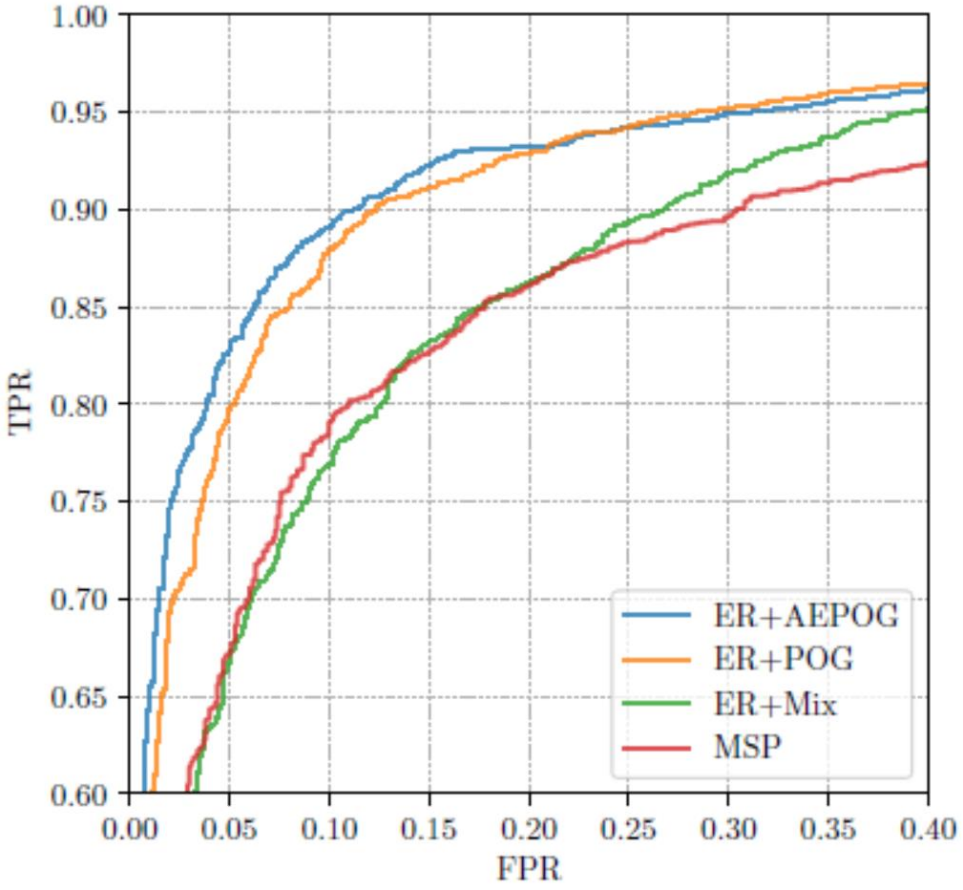
Model	AUROC↑	AUPR↑	FPR95↓	FPR90↓
Cont. GAN	69.80†	87.55†	90.77†	76.18†
Maha. Dis.	73.85†	48.20†	73.96†	61.48†
Likelihood Ratio	90.83†	96.63†	41.51†	27.37†
AE	92.41†	97.21†	37.77†	18.97†
MSP	88.41†	95.05†	52.30†	32.56†
Entropy	89.02†	95.05†	37.56†	29.31†
KNN	89.43†	95.02†	36.17†	22.91†
DOC	93.03†	97.28†	35.44†	18.07†
ODIN	90.25†	95.67†	37.92†	24.29†
ER+Perturb	96.28†	98.72*	19.23†	10.80†
ER+Mix	92.76†	97.27†	38.44†	21.71†
ER+POG	98.53*	99.42*	7.70*	4.36*
ER+AEPOG	**98.60**	**99.46**	**7.64**	**4.24**
w/o Noise	94.26†	98.07†	33.08†	18.56†
w/o Soft Token	94.08†	98.06†	40.07†	16.96†

Experiments

ROC curve on IPA dataset

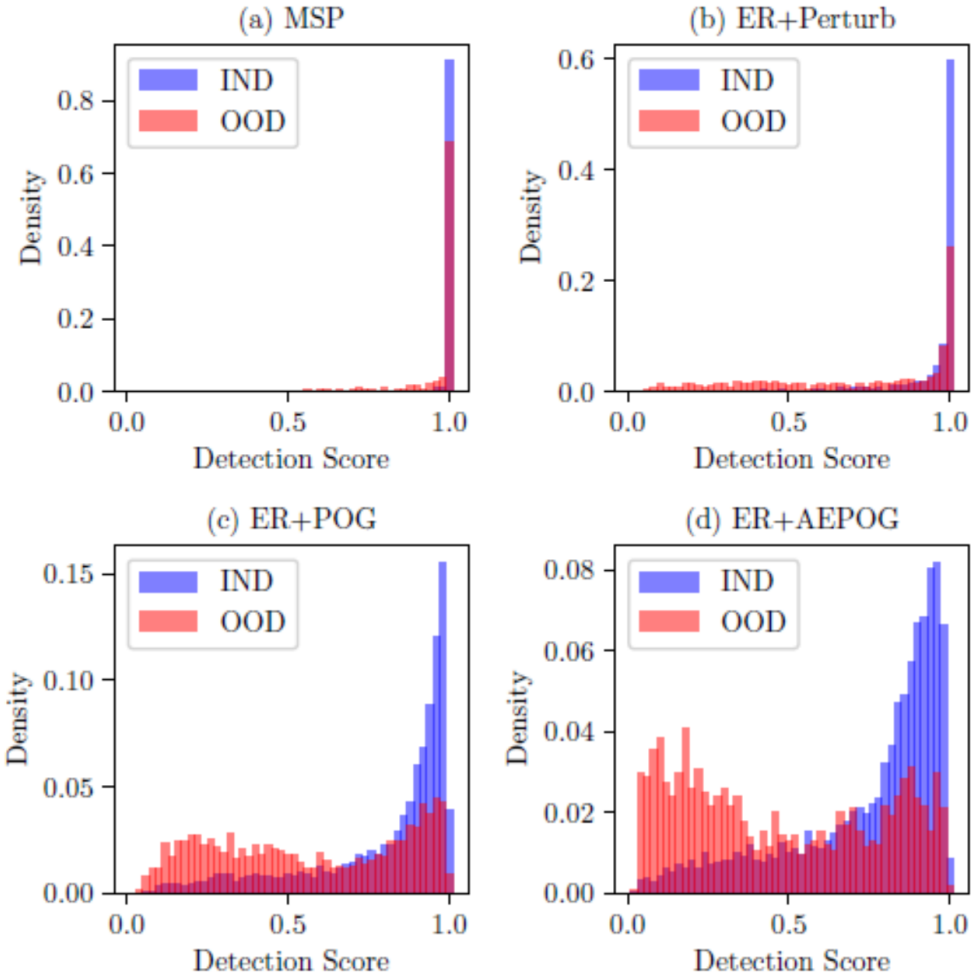


ROC curve on OSQ dataset

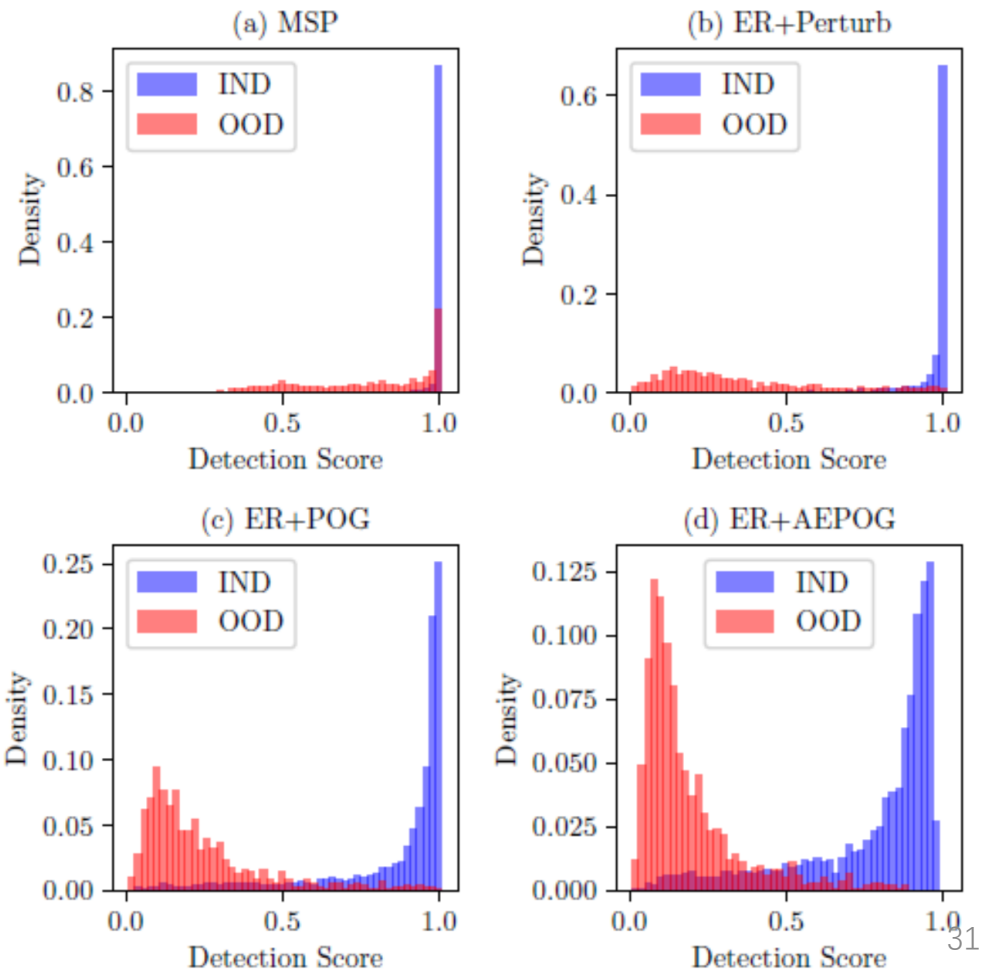


Experiments

OOD score distribution on IPA dataset

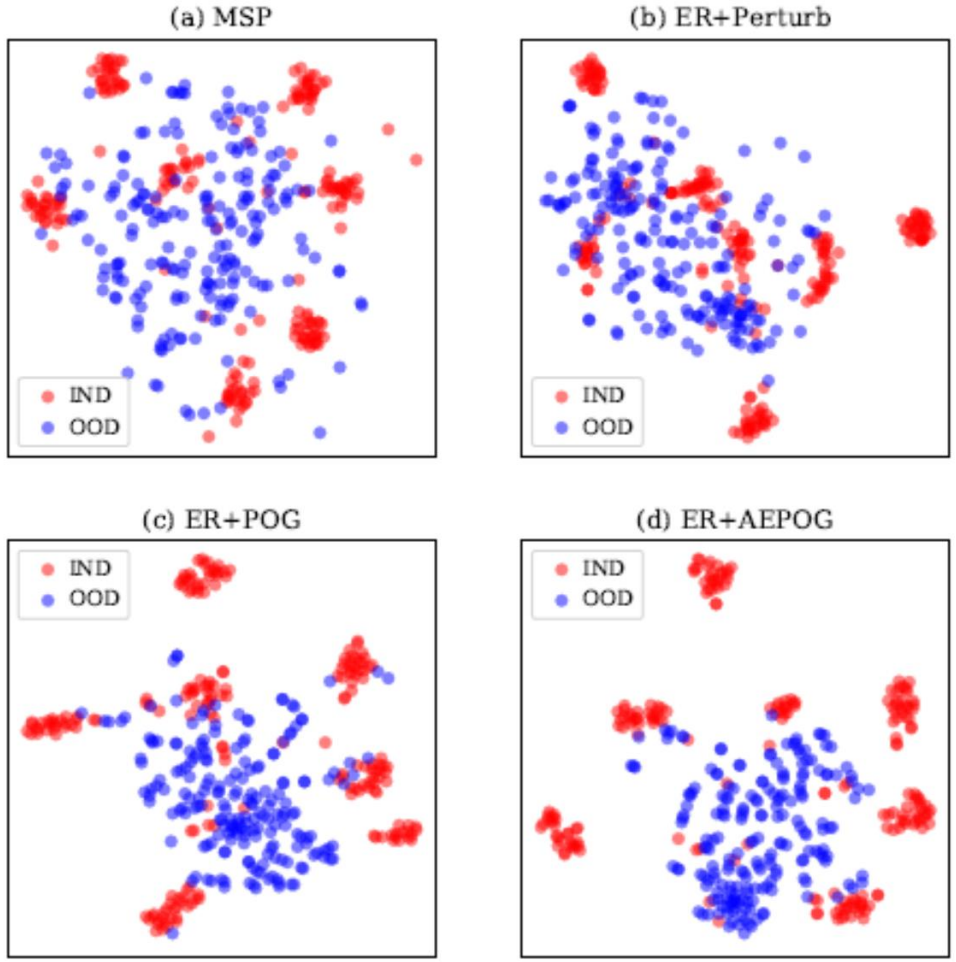


OOD score distribution on OSQ dataset



Experiments

Feature visualization on OSQ dataset



Samples of generated OOD utterances

IND Samples	Translate hello in French. Locate my phone please. Schedule a gas bill payment. Help me change my insurance plan. I'd like to improve my credit score.
OOD Samples	How much is my car worth used. Can you add a bag to my reservation. How do you fix a leaking sink. How long do wire transfer take. When was Toyota created.
Generated by POG	Please obtain French. How can make my phone get. What meetings can I schedule there. Can you help me an setting. I'd like to include the email my my dinner.
Generated by AEPOG	How much is this payments please. How good is my reservation like. How do you divided for pork. Tell me how long I taken for chilis off. When was today's name with delta vehicle.

Conclusion

- Our model can produce more effective pseudo OOD samples
- The pseudo OOD samples generated by the proposed POG model can be used to improve the OOD detection performance
- The performance on IND samples is not affected