TopicDiff: A Topic-enriched Diffusion Approach for Multimodal Conversational Emotion Detection

Jiamin Luo*, Jingjing Wang*, Guodong Zhou†
School of Computer Science and Technology, Soochow University, China
No.1, Shizi Street, Suzhou City, Jiangsu Province, China
20204027003@stu.suda.edu.cn, {djingwang, gdzhou, }@suda.edu.cn

COLING-2024

Reported by JiaWei Cheng

Introduction

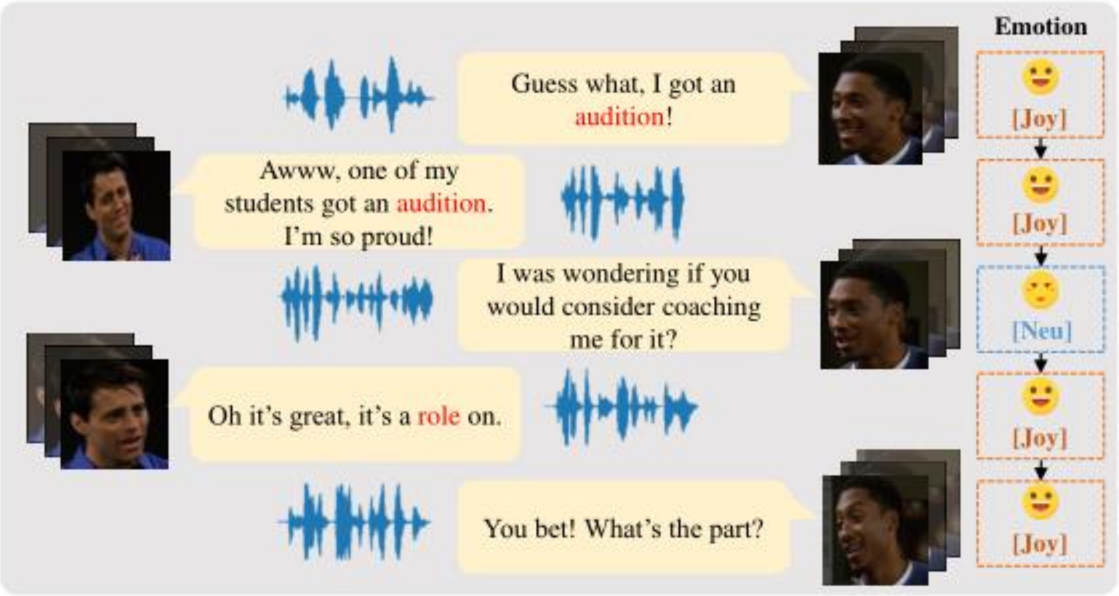
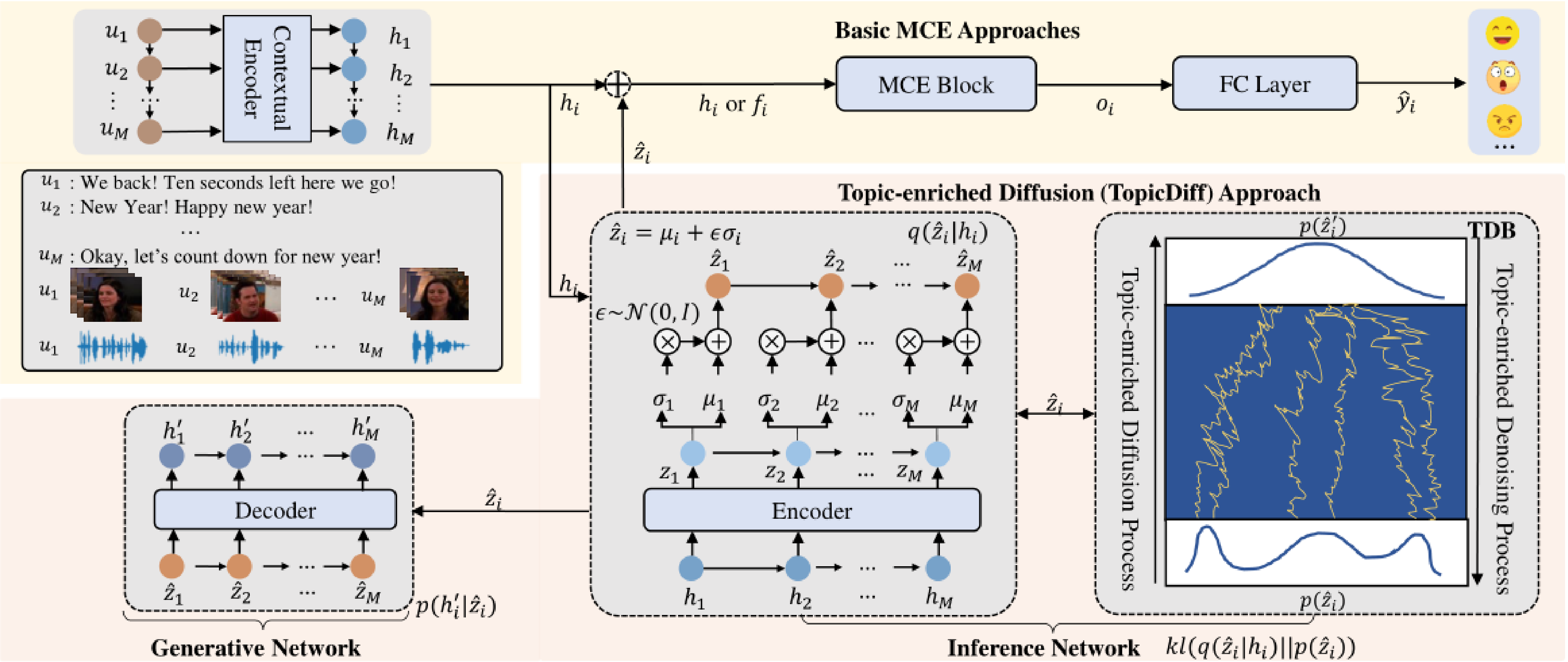

Emotion
Guess what, I got an audition!
[Joy]
Awww, one of my students got an audition. I'm so proud!
[Joy]
I was wondering if you would consider coaching me for it?
[Neu]
Oh it's great, it's a role on.
[Joy]
You bet! What's the part?
[Joy]

Figure 1: A multimodal conversational example from MELD dataset to illustrate the importance of multimodal topic information, where each utterance contains acoustic spectrum, video frame, language and corresponding emotion label.

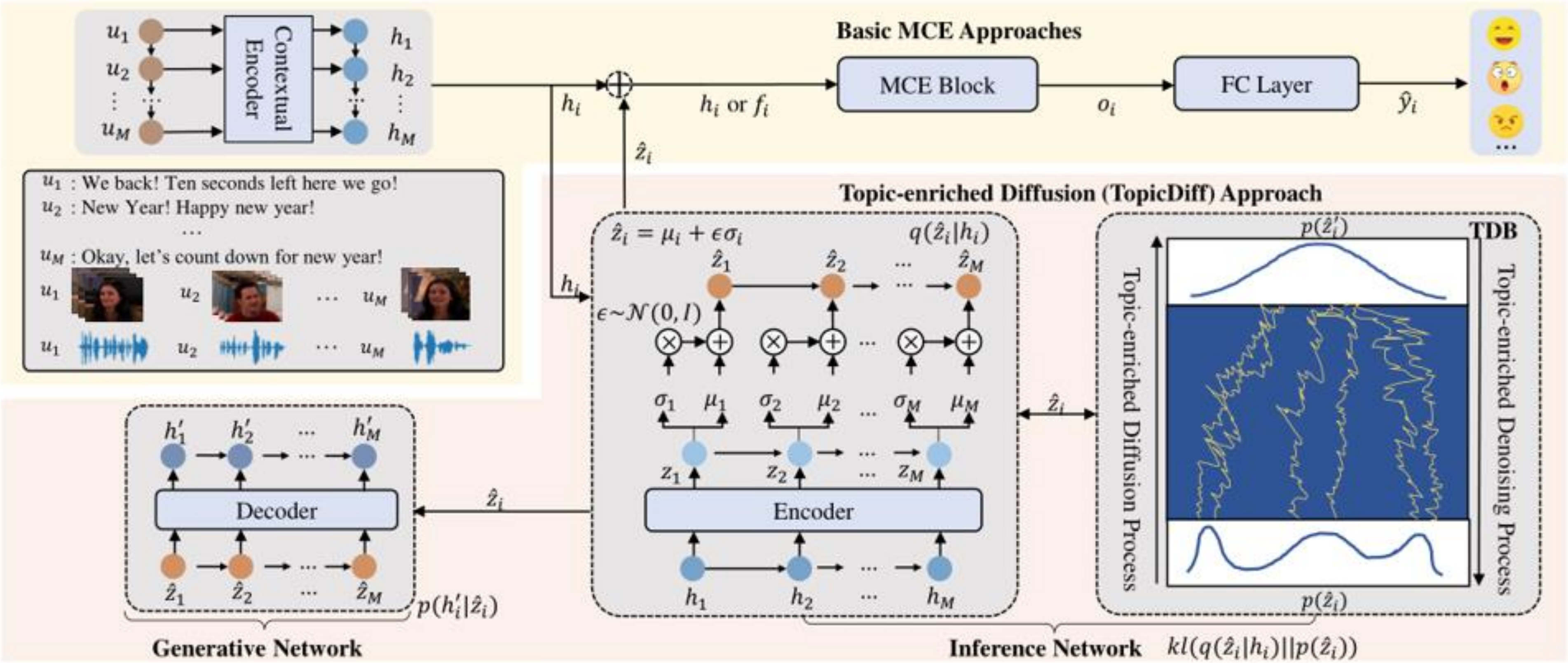
Overview


u_1
u_2
u_M
Contextual Encoder
h_1
h_2
h_M
h_i
h_i or f_i
$\hat{z}_i$
Basic MCE Approaches
MCE Block
o_i
FC Layer
$\hat{y}_i$
Topic-enriched Diffusion (TopicDiff) Approach
$\hat{z}_i = \mu_i + \epsilon\sigma_i$
$q(\hat{z}_i|h_i)$
$\hat{z}_1$ $\hat{z}_2$ ... $\hat{z}_M$
$\epsilon \sim \mathcal{N}(0, I)$
σ_1 μ_1 σ_2 μ_2 ... σ_M μ_M
z_1 z_2 ... z_M
Encoder
h_1 h_2 ... h_M
$p(\hat{z}_i')$
TDB
Topic-enriched Diffusion Process
Topic-enriched Denoising Process
$p(\hat{z}_i)$
h_1' h_2' ... h_M'
Decoder
$\hat{z}_1$ $\hat{z}_2$ ... $\hat{z}_M$
$p(h_i'|\hat{z}_i)$
Generative Network
Inference Network
$kl(q(\hat{z}_i|h_i)||p(\hat{z}_i))$

u_1 : We back! Ten seconds left here we go!
u_2 : New Year! Happy new year!
...
u_M : Okay, let's count down for new year!

Figure 2: The overall architecture of our model-agnostic Topic-enriched Diffusion (TopicDiff) approach for MCE, where TDB represents Topic-enriched Diffusion Block consisting of Topic-enriched Diffusion Process and Topic-enriched Denoising Process.

Method



u_1 : We back! Ten seconds left here we go!
u_2 : New Year! Happy new year!
...
u_M : Okay, let's count down for new year!

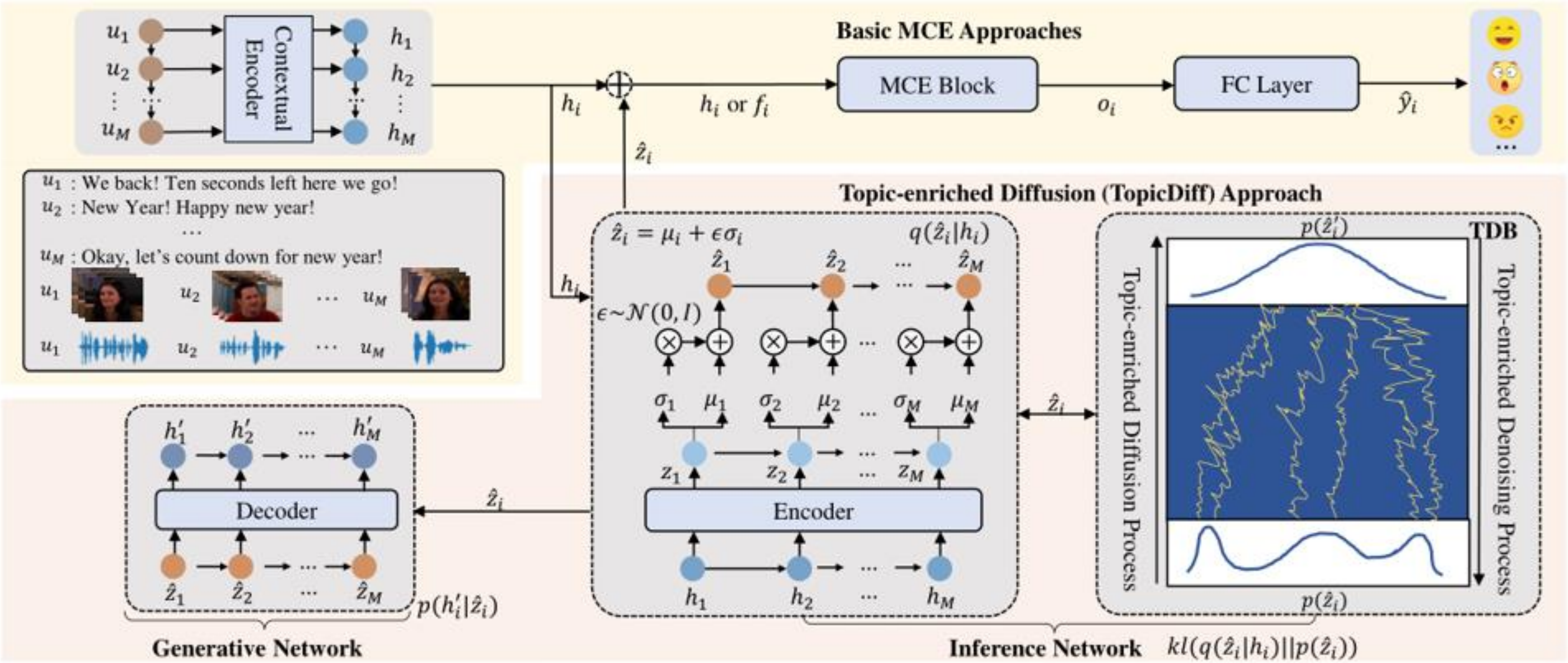
$$\boldsymbol{h}_i = \text{CTEncoder}(\boldsymbol{u}_i) \qquad (1)$$

$$\mathcal{L}_{mce} = -\frac{1}{\sum_{n=1}^{N} c(n)} \sum_{j=1}^{N} \sum_{i=1}^{c(n)} y_{j,i}^{n} \log \hat{y}_{j,i}^{n} \qquad (2)$$

$$\mathrm{d}\hat{\boldsymbol{z}}_i = f(\hat{\boldsymbol{z}}_i, t)\mathrm{d}t + g(t)\mathrm{d}\boldsymbol{w} \qquad (3)$$

$$\mathrm{d}\hat{\boldsymbol{z}}_i = [f(\hat{\boldsymbol{z}}_i, t) - g(t)^2 \nabla_{\hat{\boldsymbol{z}}_i} \log p(\hat{\boldsymbol{z}}_i)]\mathrm{d}t + g(t)\mathrm{d}\hat{\boldsymbol{w}} \qquad (4)$$

Method

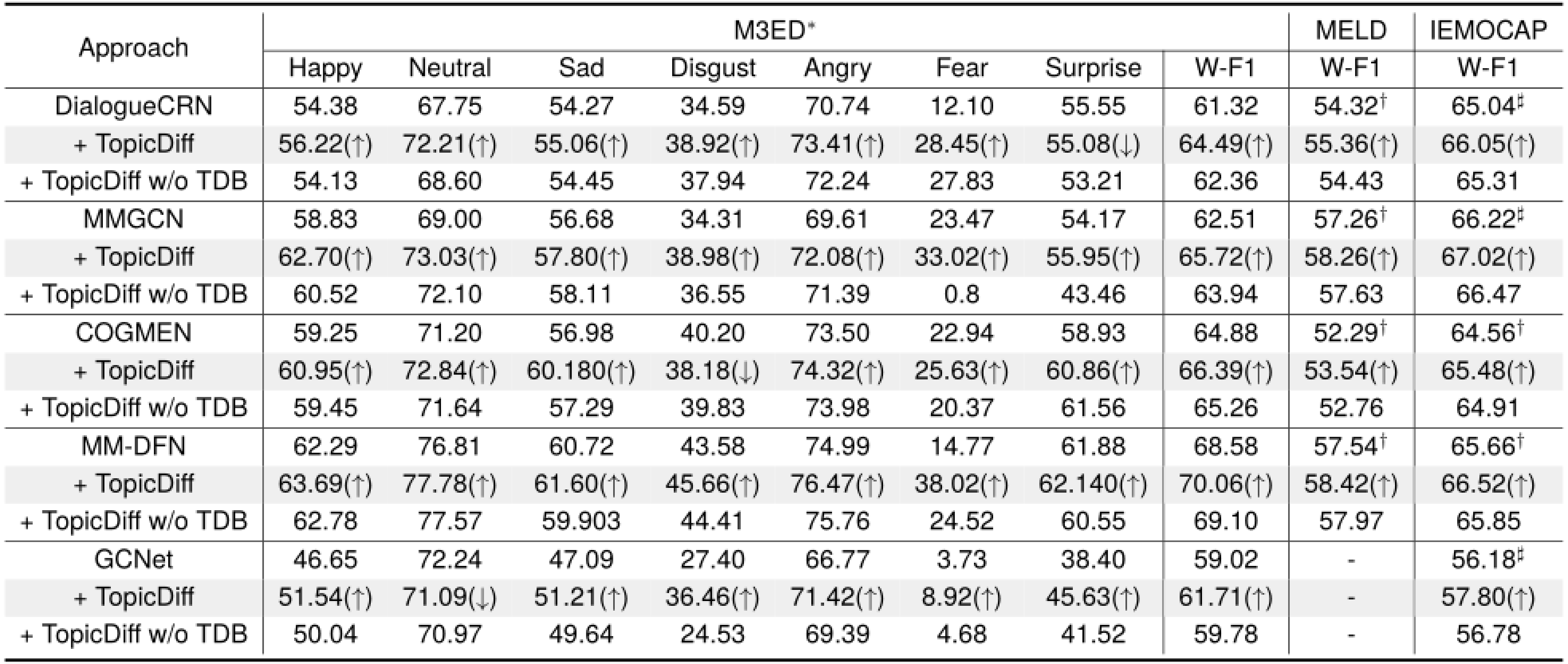


$$\mathcal{L}_{rec} = \mathbb{E}_{q(\hat{\boldsymbol{z}}_i|\boldsymbol{h}_i)}\left[\log p(\boldsymbol{h}_i^{'}|\hat{\boldsymbol{z}}_i)\right] \qquad (5)$$

$$\mathcal{L}_{kl} = kl(q(\hat{\boldsymbol{z}}_i|\boldsymbol{h}_i)||p(\hat{\boldsymbol{z}}_i)) \qquad (6)$$

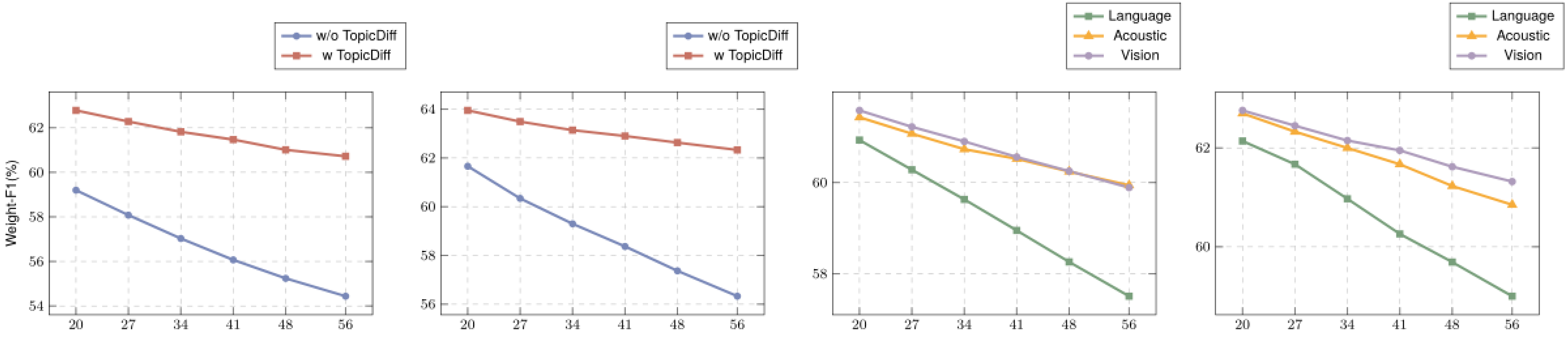
$$\mathcal{L}_{total}=\mathcal{L}_{mce}+\alpha\sum\nolimits_{(\boldsymbol{a},\boldsymbol{v},\boldsymbol{l})}\mathcal{L}_{rec}+\beta\sum\nolimits_{(\boldsymbol{a},\boldsymbol{v},\boldsymbol{l})}\mathcal{L}_{kl} \qquad (7)$$

Experiments



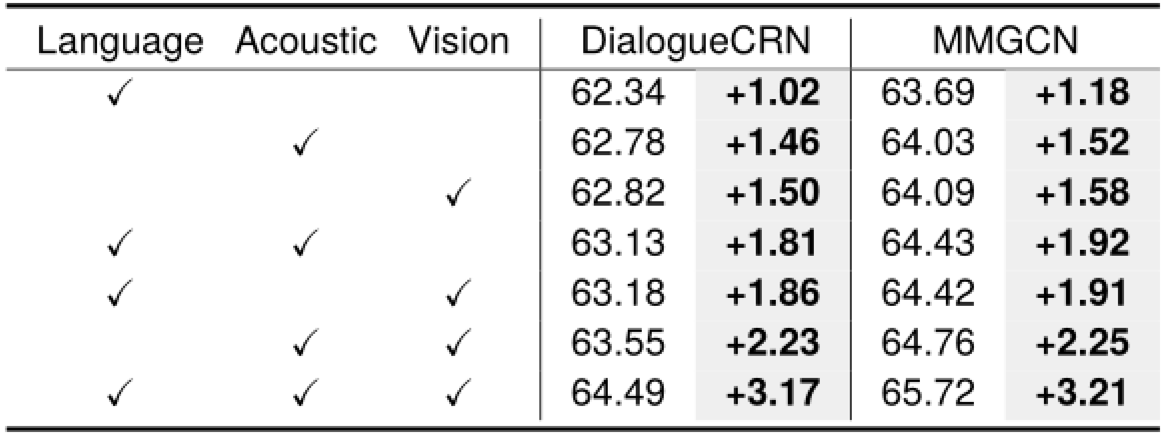
Approach	M3ED*								MELD	IEMOCAP
	Happy	Neutral	Sad	Disgust	Angry	Fear	Surprise	W-F1	W-F1	W-F1
DialogueCRN	54.38	67.75	54.27	34.59	70.74	12.10	55.55	61.32	54.32†	65.04$^{\sharp}$
+ TopicDiff	56.22(↑)	72.21(↑)	55.06(↑)	38.92(↑)	73.41(↑)	28.45(↑)	55.08(↓)	64.49(↑)	55.36(↑)	66.05(↑)
+ TopicDiff w/o TDB	54.13	68.60	54.45	37.94	72.24	27.83	53.21	62.36	54.43	65.31
MMGCN	58.83	69.00	56.68	34.31	69.61	23.47	54.17	62.51	57.26†	66.22$^{\sharp}$
+ TopicDiff	62.70(↑)	73.03(↑)	57.80(↑)	38.98(↑)	72.08(↑)	33.02(↑)	55.95(↑)	65.72(↑)	58.26(↑)	67.02(↑)
+ TopicDiff w/o TDB	60.52	72.10	58.11	36.55	71.39	0.8	43.46	63.94	57.63	66.47
COGMEN	59.25	71.20	56.98	40.20	73.50	22.94	58.93	64.88	52.29†	64.56†
+ TopicDiff	60.95(↑)	72.84(↑)	60.180(↑)	38.18(↓)	74.32(↑)	25.63(↑)	60.86(↑)	66.39(↑)	53.54(↑)	65.48(↑)
+ TopicDiff w/o TDB	59.45	71.64	57.29	39.83	73.98	20.37	61.56	65.26	52.76	64.91
MM-DFN	62.29	76.81	60.72	43.58	74.99	14.77	61.88	68.58	57.54†	65.66†
+ TopicDiff	63.69(↑)	77.78(↑)	61.60(↑)	45.66(↑)	76.47(↑)	38.02(↑)	62.140(↑)	70.06(↑)	58.42(↑)	66.52(↑)
+ TopicDiff w/o TDB	62.78	77.57	59.903	44.41	75.76	24.52	60.55	69.10	57.97	65.85
GCNet	46.65	72.24	47.09	27.40	66.77	3.73	38.40	59.02	-	56.18$^{\sharp}$
+ TopicDiff	51.54(↑)	71.09(↓)	51.21(↑)	36.46(↑)	71.42(↑)	8.92(↑)	45.63(↑)	61.71(↑)	-	57.80(↑)
+ TopicDiff w/o TDB	50.04	70.97	49.64	24.53	69.39	4.68	41.52	59.78	-	56.78

Experiments



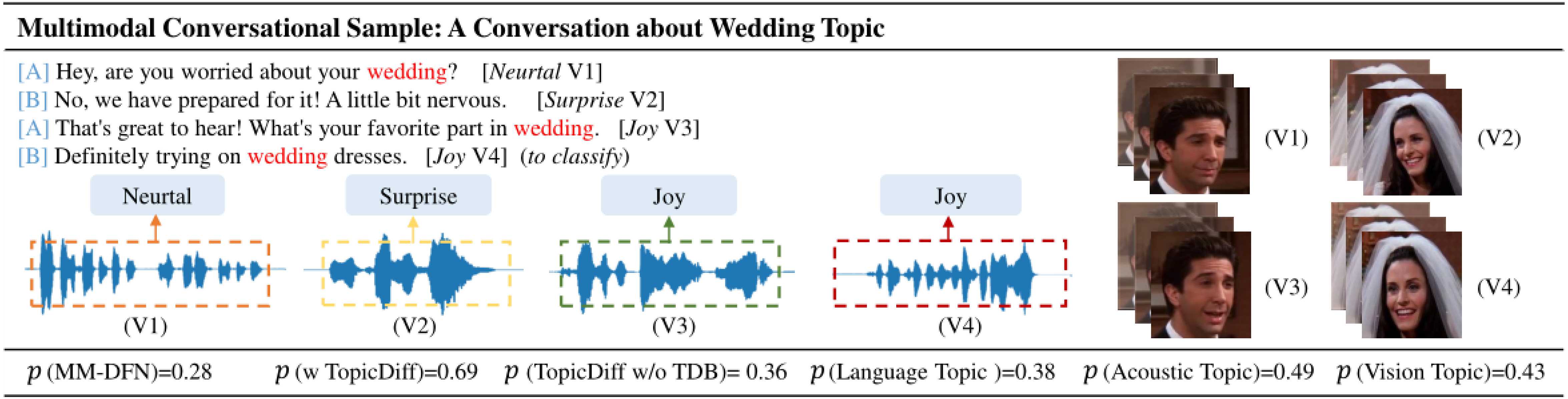
(a) DialogueCRN (b) MMGCN (c) DialogueCRN+TopicDiff (d) MMGCN+TopicDiff

Experiments



Language	Acoustic	Vision	DialogueCRN		MMGCN	
✓			62.34	**+1.02**	63.69	**+1.18**
	✓		62.78	**+1.46**	64.03	**+1.52**
		✓	62.82	**+1.50**	64.09	**+1.58**
✓	✓		63.13	**+1.81**	64.43	**+1.92**
✓		✓	63.18	**+1.86**	64.42	**+1.91**
	✓	✓	63.55	**+2.23**	64.76	**+2.25**
✓	✓	✓	64.49	**+3.17**	65.72	**+3.21**

Experiments



Multimodal Conversational Sample: A Conversation about Wedding Topic

[A] Hey, are you worried about your wedding? [*Neurtal* V1]
[B] No, we have prepared for it! A little bit nervous. [*Surprise* V2]
[A] That's great to hear! What's your favorite part in wedding. [*Joy* V3]
[B] Definitely trying on wedding dresses. [*Joy* V4] (*to classify*)

p (MM-DFN)=0.28 p (w TopicDiff)=0.69 p (TopicDiff w/o TDB)= 0.36 p (Language Topic)=0.38 p (Acoustic Topic)=0.49 p (Vision Topic)=0.43

Thanks！