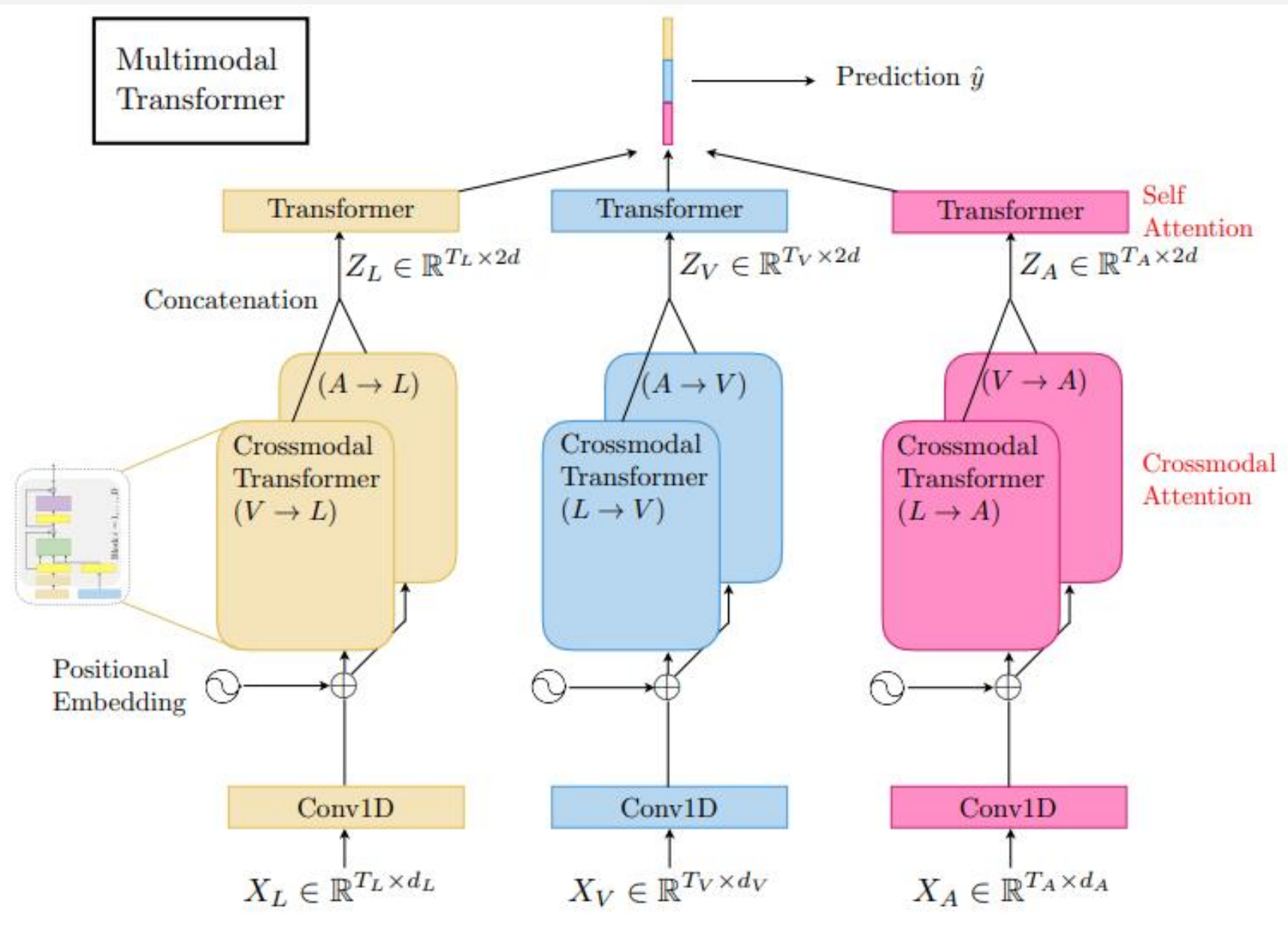
Efficient Multimodal Transformer with Dual-Level Feature Restoration for Robust Multimodal Sentiment Analysis

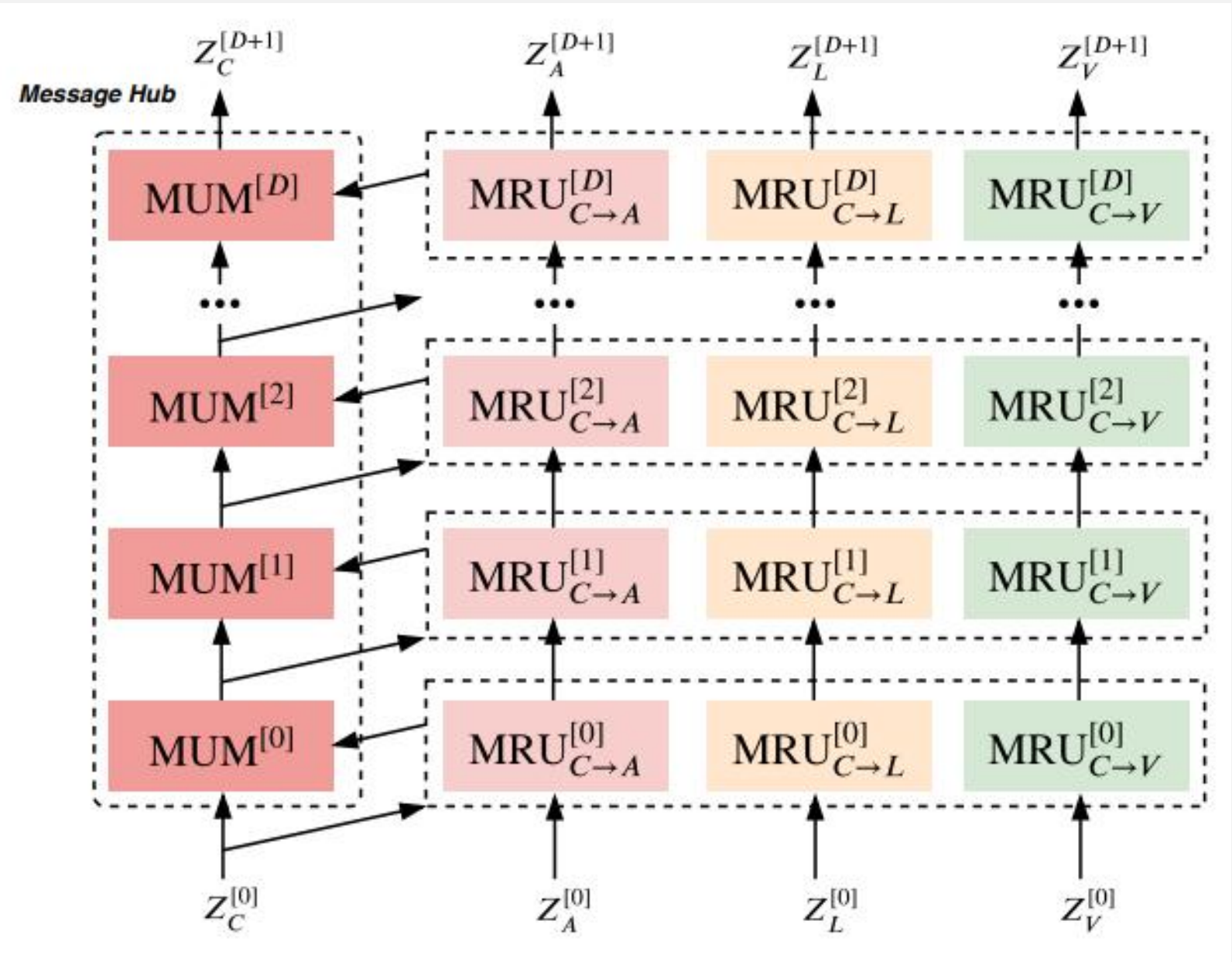
Licai Sun[1], Zheng Lian[2], Bin Liu[2], Jianhua Tao[3]

[1]UCAS & [2]CASIA & [3] Tsinghua University

How to fuse unaligned multimodal sequence?

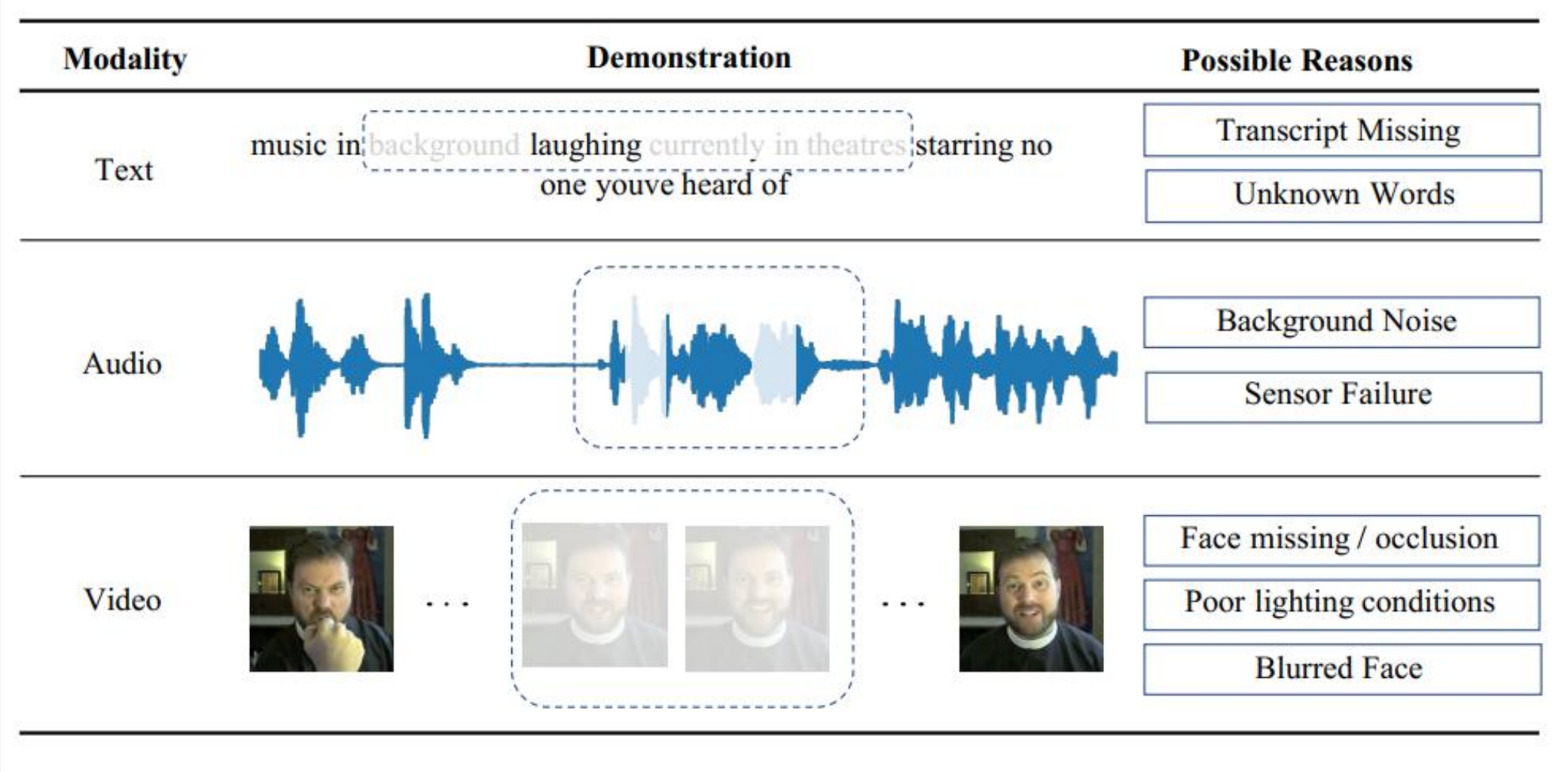

Multimodal Transformer
Prediction $\hat{y}$
Transformer
Self Attention
$Z_L \in \mathbb{R}^{T_L \times 2d}$
$Z_V \in \mathbb{R}^{T_V \times 2d}$
$Z_A \in \mathbb{R}^{T_A \times 2d}$
Concatenation
Crossmodal Transformer $(V \to L)$
Crossmodal Transformer $(L \to V)$
Crossmodal Transformer $(L \to A)$
Crossmodal Attention
Positional Embedding
Conv1D
$X_L \in \mathbb{R}^{T_L \times d_L}$
$X_V \in \mathbb{R}^{T_V \times d_V}$
$X_A \in \mathbb{R}^{T_A \times d_A}$

MulT (ACL 2019)
One-to-One Local-Local Fusion
Complexity: O(M²T²)

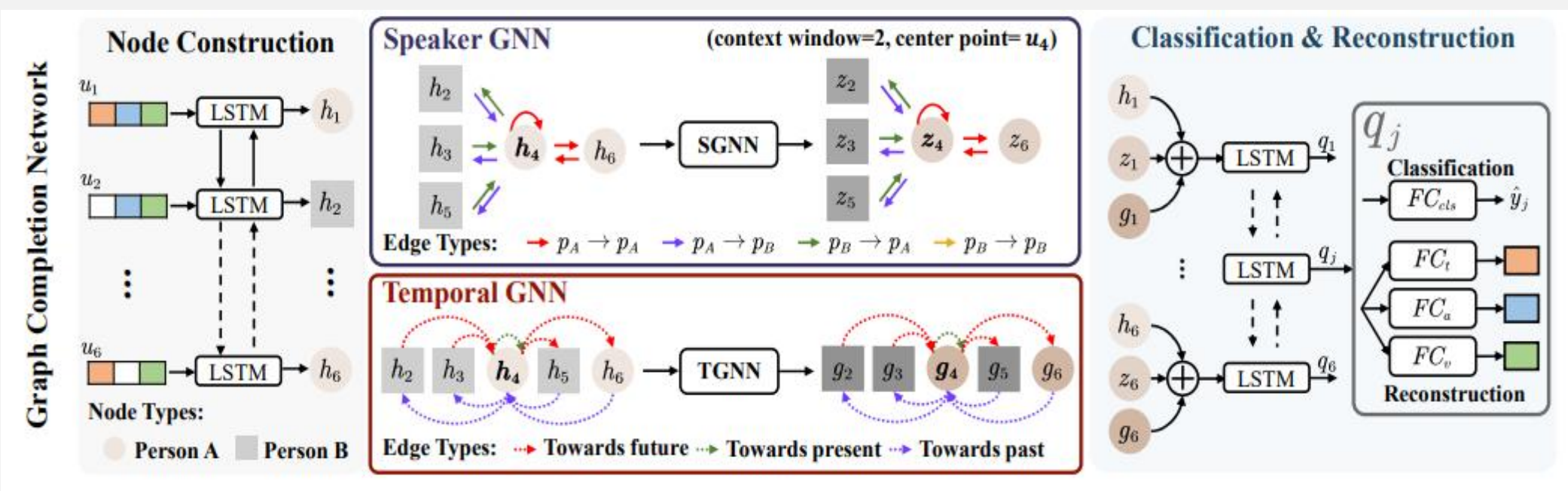


PMR (CVPR 2021)
One-to-All Local-Local Fusion
Complexity: O(M²T²)

How to handle random feature missing?


Modality
Demonstration
Possible Reasons
Text
music in background laughing currently in theaters starring no one youve heard of
Transcript Missing
Unknown Words
Audio
Background Noise
Sensor Failure
Video
Face missing / occlusion
Poor lighting conditions
Blurred Face

TFR-Net (MM 2021)
Low-Level Feature Reconstruction



GCNet (TPAMI 2023)
Low-Level Feature Reconstruction

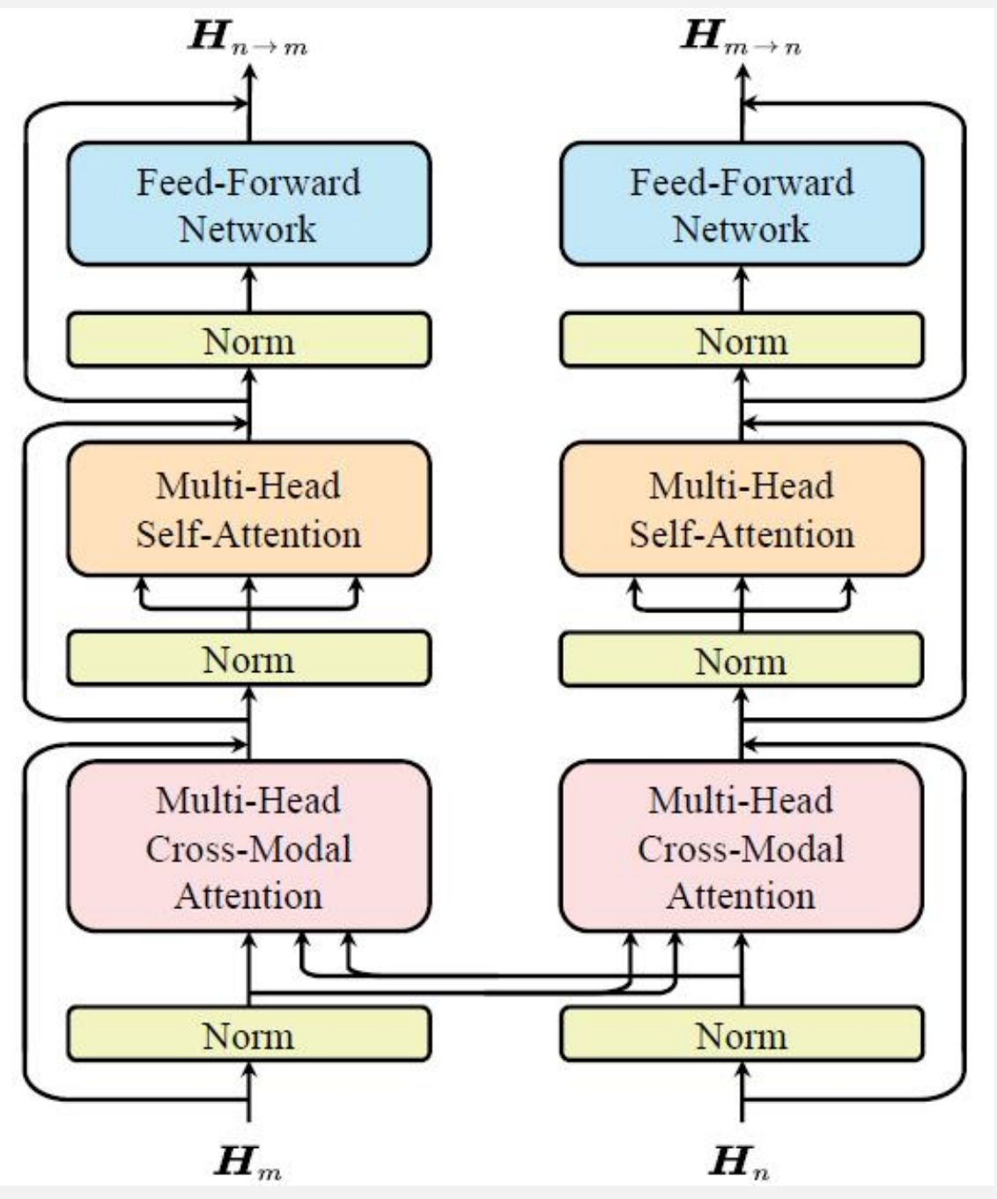
Our solution: EMT-DLFR

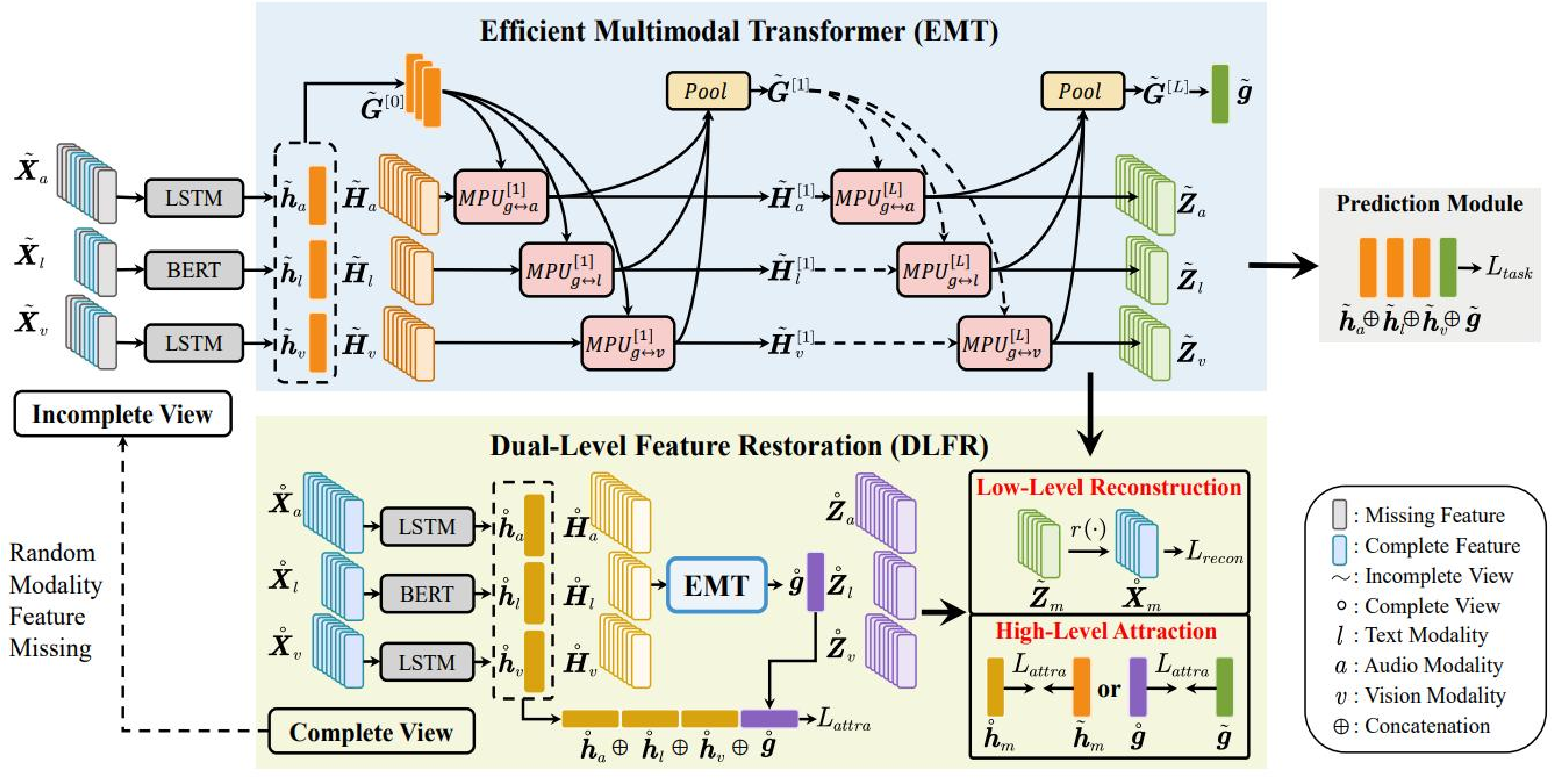
EMT

One-to-All Global-Local Fusion
Complexity: O(MT²)

$$H_l^{[i+1]}, G_{l\to g}^{[i]} = \mathrm{MPU}_{l\leftrightarrow g}^{[i]}(H_l^{[i]}, G^{[i]})$$
$$H_a^{[i+1]}, G_{a\to g}^{[i]} = \mathrm{MPU}_{a\leftrightarrow g}^{[i]}(H_a^{[i]}, G^{[i]})$$
$$H_v^{[i+1]}, G_{v\to g}^{[i]} = \mathrm{MPU}_{v\leftrightarrow g}^{[i]}(H_v^{[i]}, G^{[i]})$$

MPU

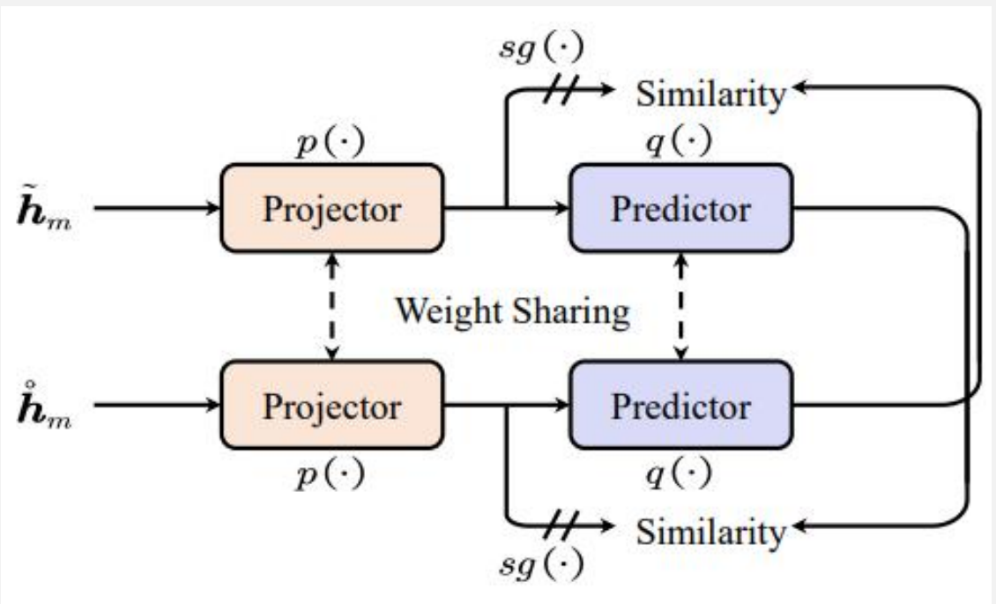

$\boldsymbol{H}_{n\to m}$
$\boldsymbol{H}_{m\to n}$
Feed-Forward Network
Norm
Multi-Head Self-Attention
Multi-Head Cross-Modal Attention
$\boldsymbol{H}_m$
$\boldsymbol{H}_n$



DLFR

Low-Level Feature Reconstruction (LLFR)
+
High-Level Feature Attraction (HLFR)

$$\mathcal{L}^l_{\text{recon}} = \text{smooth}_{\text{L1}}((\mathring{H}_l - r(\tilde{Z}_l)) \cdot (1 - g_l))$$
$$\mathcal{L}^a_{\text{recon}} = \text{smooth}_{\text{L1}}((\mathring{X}_a - r(\tilde{Z}_a)) \cdot (1 - g_a))$$
$$\mathcal{L}^v_{\text{recon}} = \text{smooth}_{\text{L1}}((\mathring{X}_v - r(\tilde{Z}_v)) \cdot (1 - g_v))$$
$$\mathcal{L}_{\text{recon}} = \sum_{m\in\{l,a,v\}} \mathcal{L}^m_{\text{recon}}$$



SimSiam for HLFR

Quantitative results

Complete setting

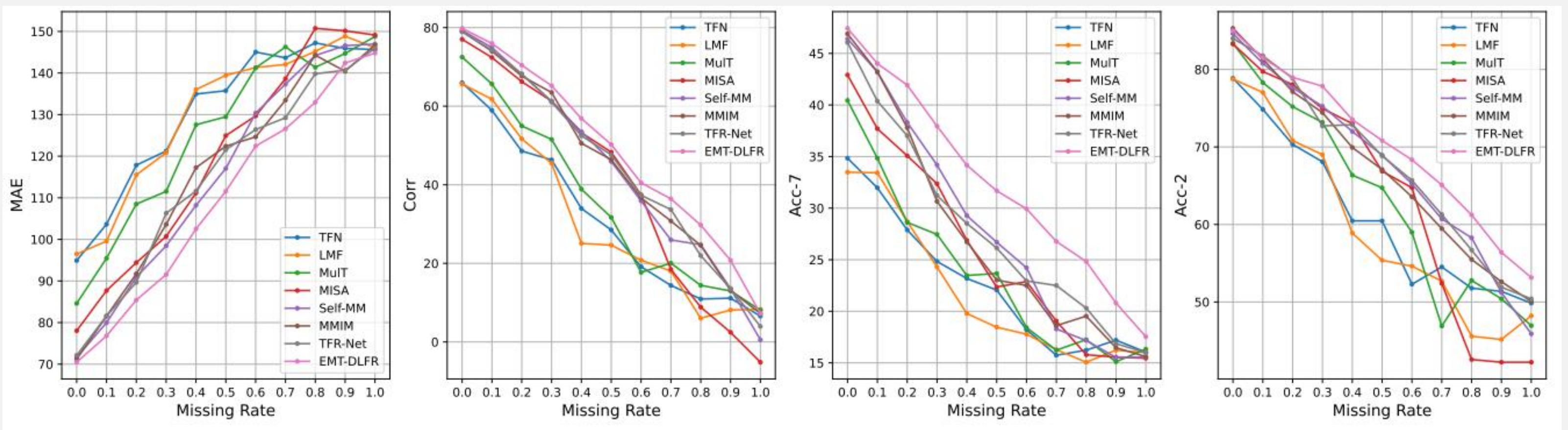
Models	MAE (↓)	Corr (↑)	Acc-7 (↑)	Acc-5 (↑)	Acc-2 (↑)	F1 (↑)
	CMU-MOSI					
TFN†	0.901	0.698	34.9	-	-/80.8	-/80.7
LMF†	0.917	0.695	33.2	-	-/82.5	-/82.4
MulT†	0.861	0.711	-	-	81.5/84.1	80.6/83.9
MISA†	0.804	0.764	-	-	80.8/82.1	80.8/82.0
Self-MM†	0.712	0.795	45.8	-	82.5/84.8	82.7/84.9
MMIM†	0.700	0.800	46.7	-	84.1/86.1	84.0/86.0
AMML‡	0.723	0.792	46.3	-	-/84.9	-/84.8
TFR-Net◇	0.754	0.783	-	54.7	-/84.1	-/-
MulT	0.846	0.725	40.4	46.7	81.7/83.4	81.9/83.5
Self-MM	0.717	0.793	46.4	52.8	82.9/84.6	82.8/84.6
MMIM	0.712	0.790	46.9	53.0	**83.3/85.3**	**83.4/85.4**
TFR-Net	0.721	0.789	46.1	53.2	82.7/84.0	82.7/84.0
EMT	**0.705**	**0.798**	**47.4**	**54.1**	**83.3**/85.0	83.2/85.0

EMT achieves on-par/better performance

Incomplete setting

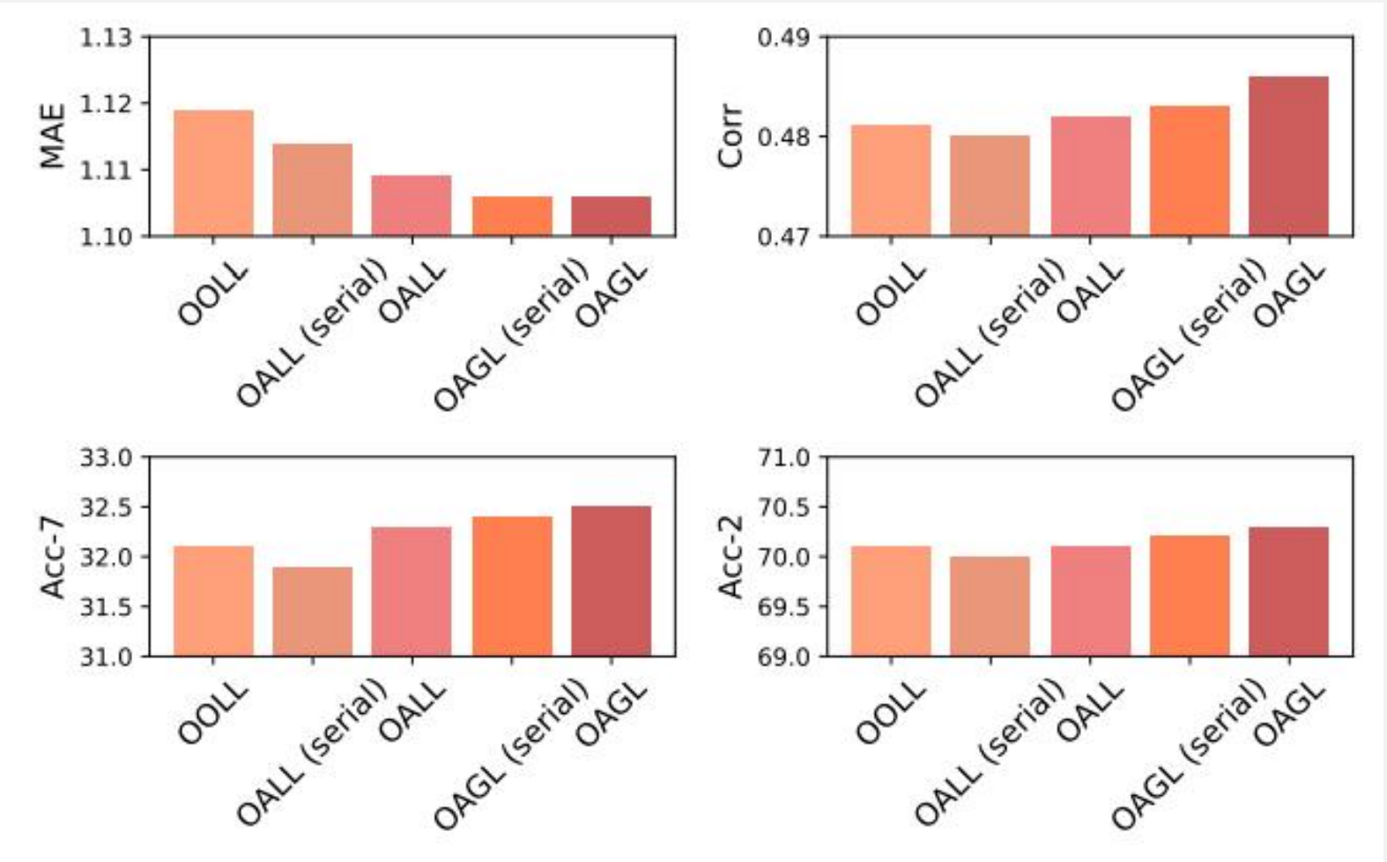
Models	MAE (↓)	Corr (↑)	Acc-7 (↑)	Acc-5 (↑)	Acc-2 (↑)	F1 (↑)
	CMU-MOSI					
TFN◇	1.327	0.300	-	23.3	-/60.4	-/-
MulT◇	1.288	0.334	-	24.4	-/61.8	-/-
MISA◇	1.209	0.403	-	27.1	-/63.2	-/-
TFR-Net◇	1.155	0.467	-	30.4	-/69.0	-/-
TFN	1.316	0.308	22.3	23.7	61.0/60.9	59.7/59.7
LMF	1.310	0.299	21.5	22.7	59.7/59.3	56.4/56.1
MulT	1.263	0.348	23.1	24.6	63.1/63.2	60.7/61.0
MISA	1.202	0.405	25.7	27.4	63.9/63.7	59.0/58.8
Self-MM	1.162	0.444	27.8	30.3	66.9/67.5	65.4/66.2
MMIM	1.168	0.450	27.0	29.4	66.8/66.9	64.6/65.8
TFR-Net	1.156	0.452	27.7	30.5	67.6/67.8	65.7/66.1
EMT-DLFR	**1.106**	**0.486**	**32.5**	**35.6**	**69.6/70.3**	**69.6/70.3**

EMT-DLFR achieves much better performance



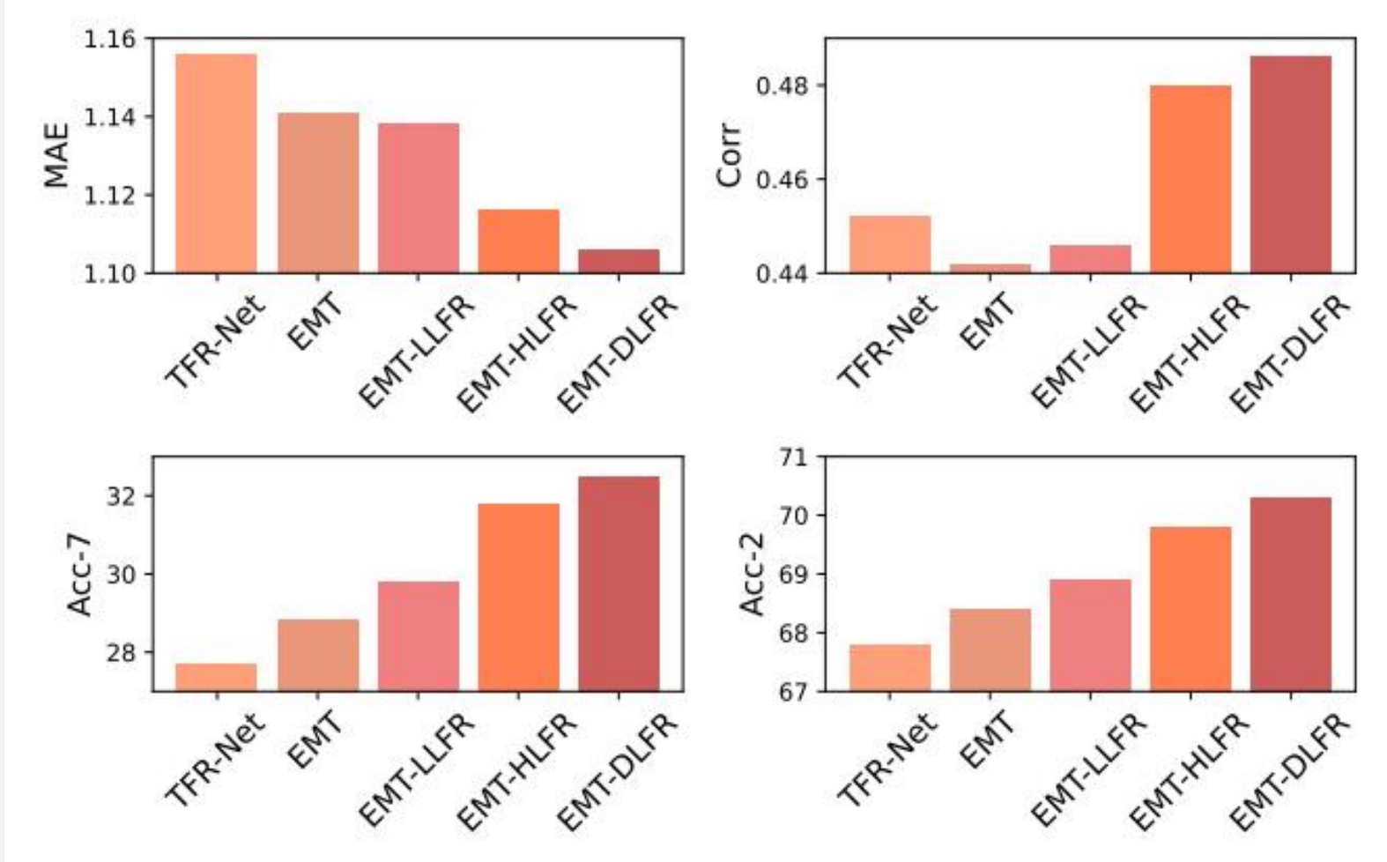
Ablation study

Fusion Strategy	MACs (G)	#Params (M)	Training Time (s)	GPU Memory (GB)
MulT	-	111.0	17.5	17.8
TFR-Net	-	124.3	24.8	16.9
OOLL	3.1	**110.5**	17.1	17.8
OALL	8.3	**110.5**	21.2	31.5
OAGL	**1.5**	**110.5**	**15.4**	**10.8**

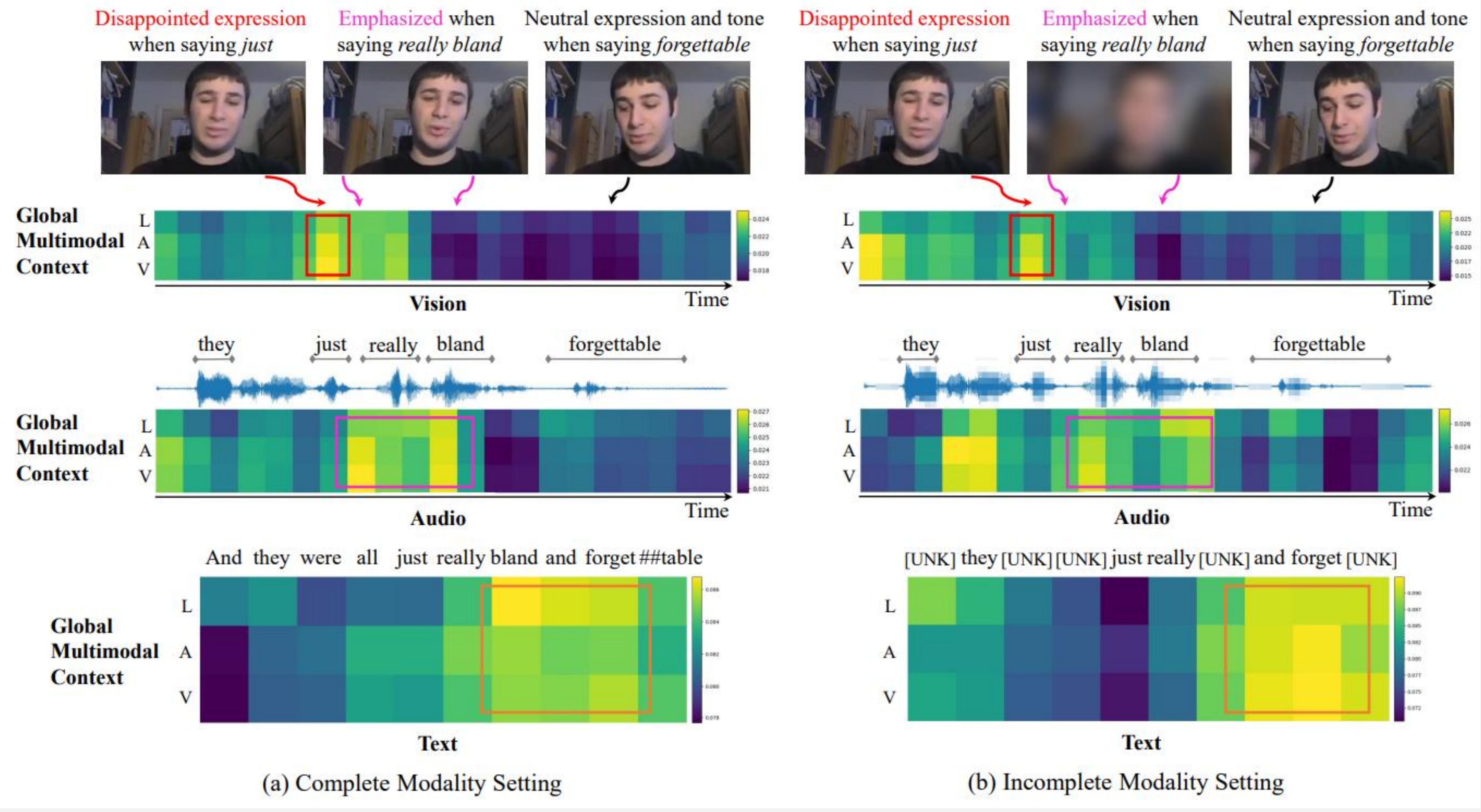


- **EMT** is *effective* and *efficient*!

- **HLFR** is more *effective* than LLFR and they are *complementary*!



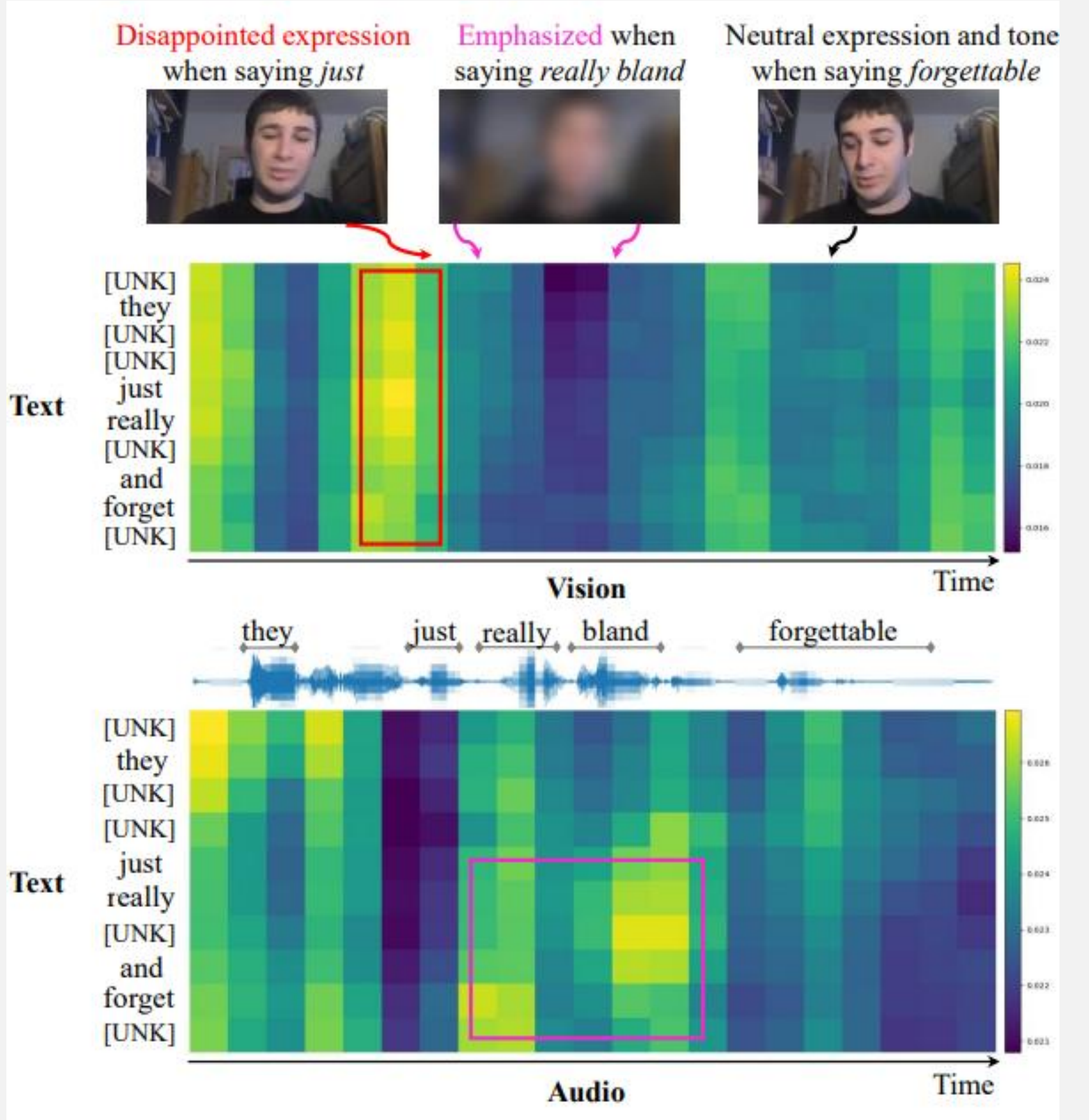
Qualitative results



(a) Complete Modality Setting

(b) Incomplete Modality Setting

- **EMT-DLFR** is *robust* to random feature missing!



- Previous local-local fusion is *low-rank* and *redundant*!

More information can be found in our paper and code:

Paper: Code: