NEURAL INFORMATION PROCESSING SYSTEMS

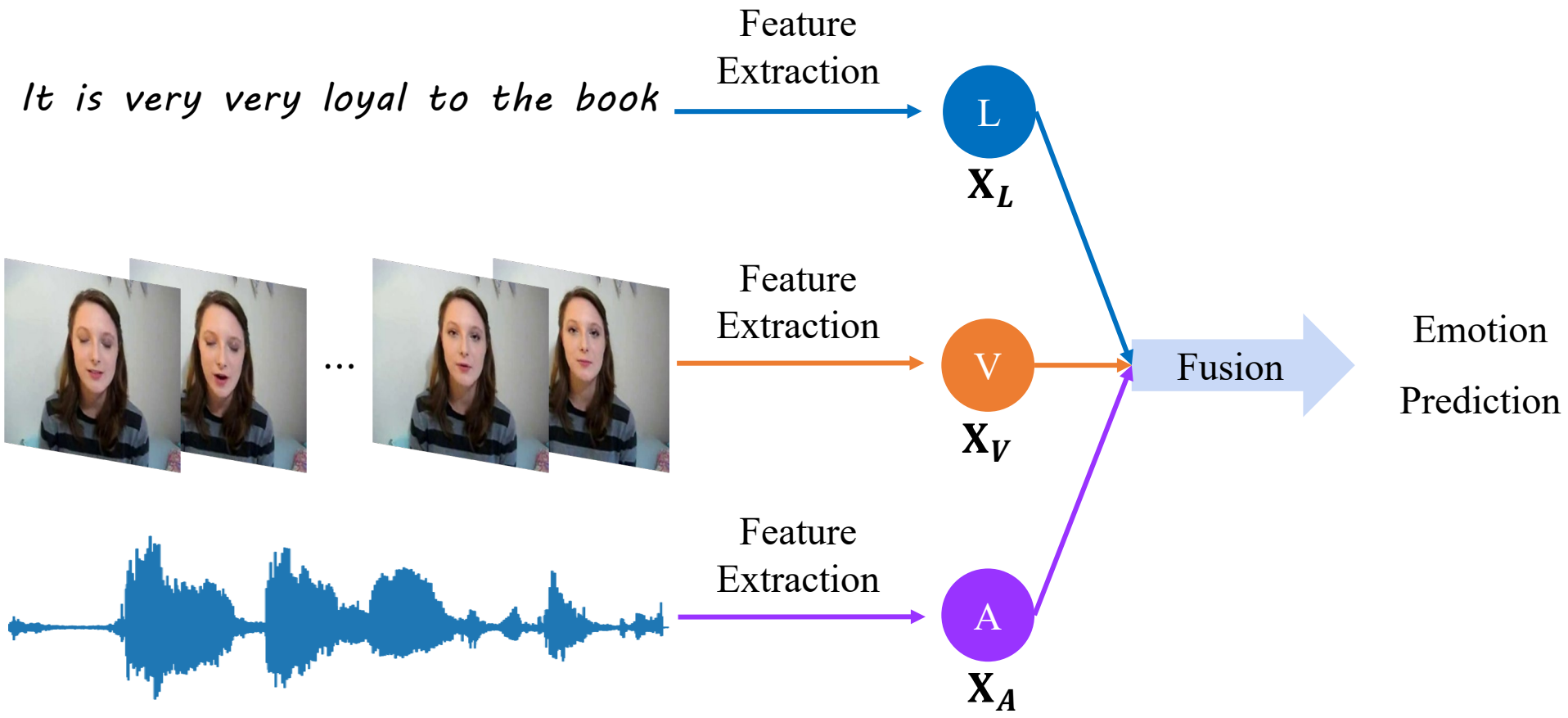
Incomplete Multimodality-Diffused Emotion Recognition

Yuanzhi Wang, Yong Li, Zhen Cui
Nanjing University of Science and Technology, Nanjing, China

Reporter： **Yuanzhi Wang**

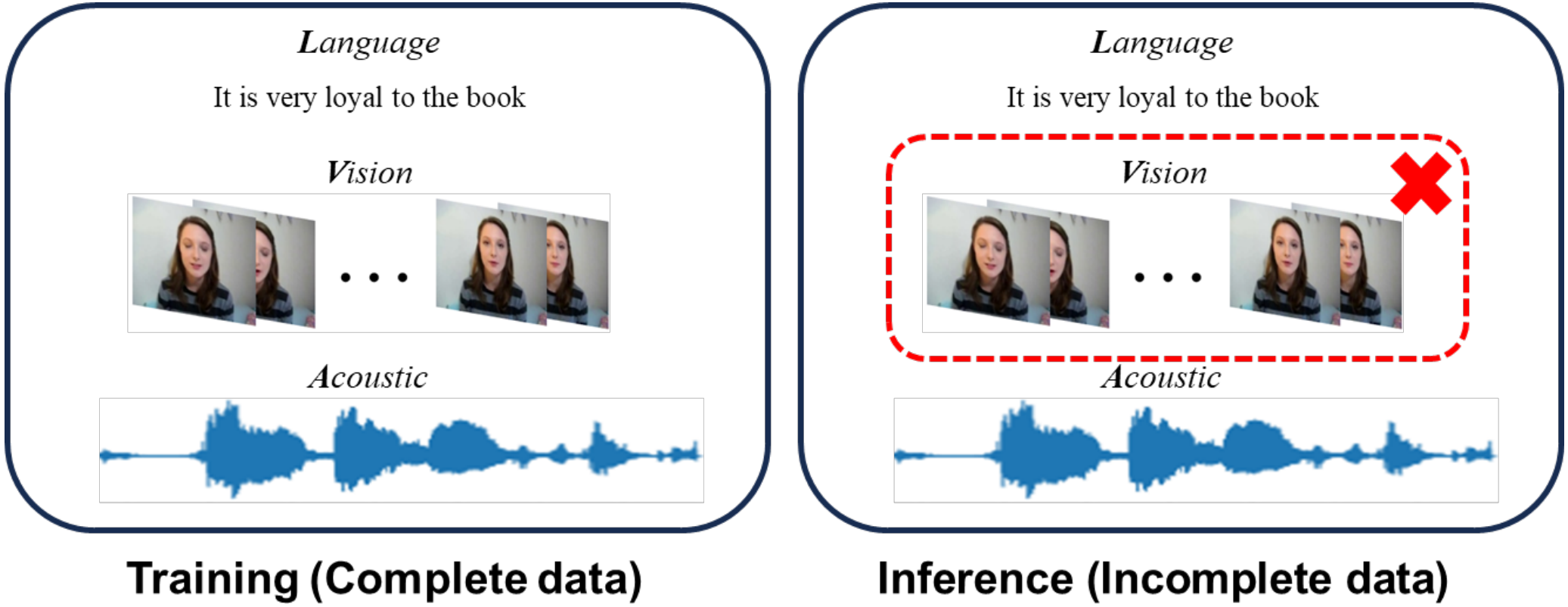
Background

- Human **Multimodal Emotion Recognition (MER)** aims to perceive the emotion of humans from video clips.
- Video clips involve multimodal temporal data, such as natural **language**, **visual** actions and **acoustic** behaviors.



Motivation

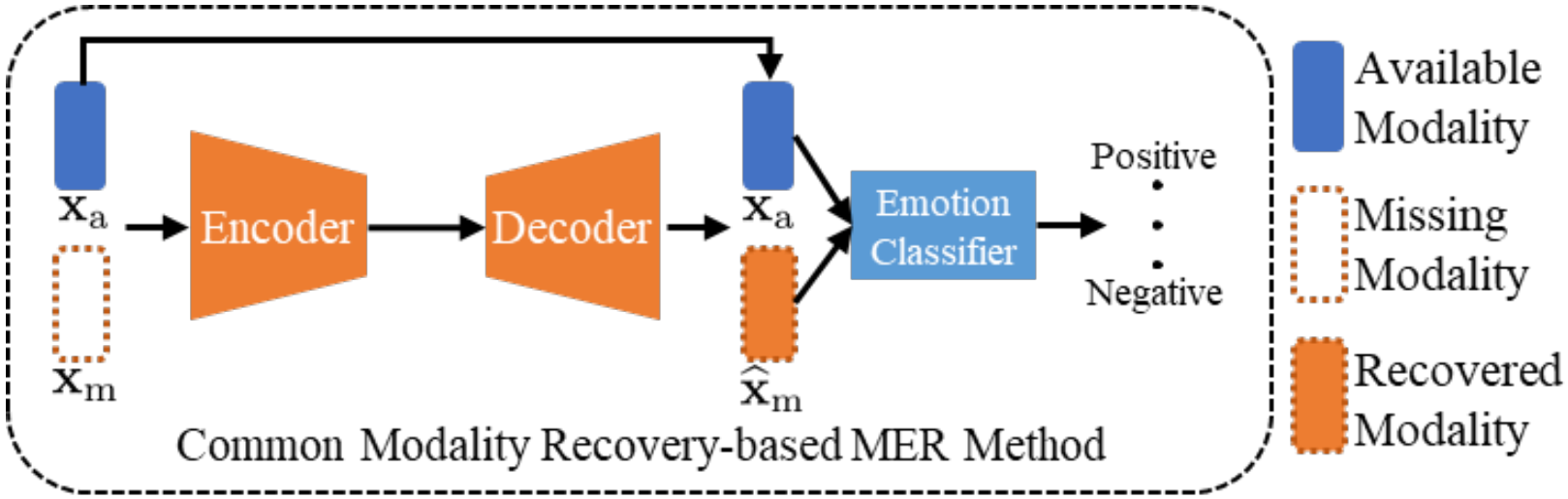
- In real-world scenarios, the well-trained MER model may be deployed when certain **modalities are missing** such as below:



A missing modality sample. The vision modality maybe missing due to social privacy security.

Motivation

- Recovering the missing modalities directly from the available ones by the well-crafted encoder and decoder often fails to explicitly consider the modality specific distributions that are highly correlated with each modality's intrinsic discriminability.

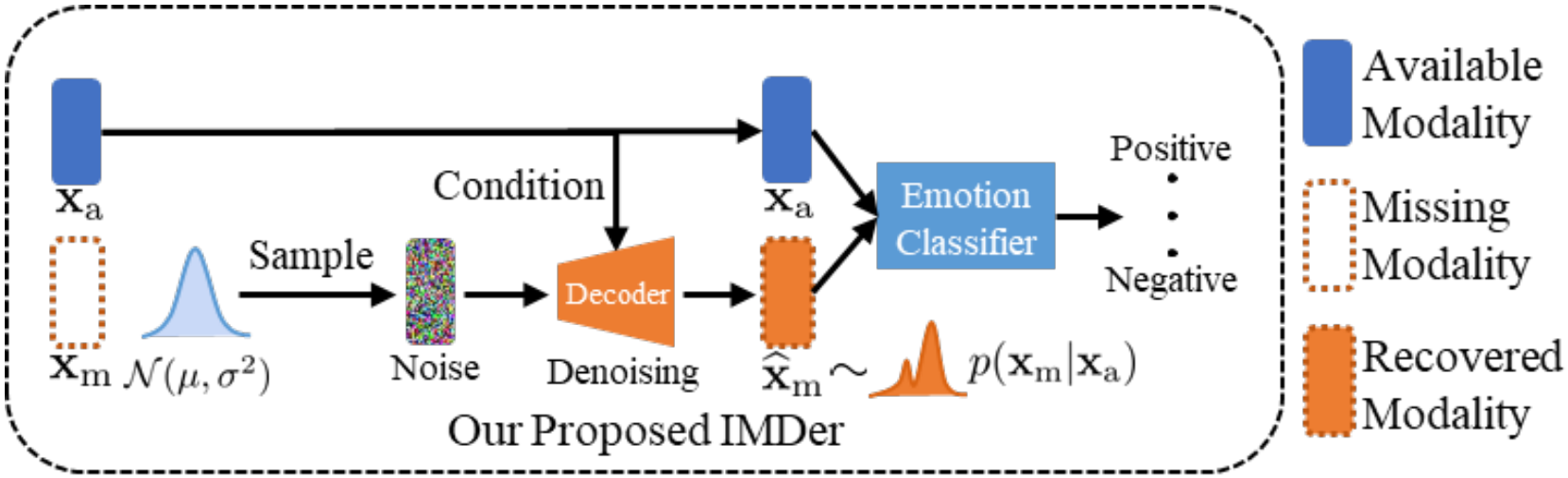


A common modality recovery method, such as [a][b].

[a] Gcnet: Graph completion network for incomplete multimodal learning in conversation. IEEE TPAMI, 2023.
[b] Found in translation: Learning robust joint representations by cyclic translations between modalities. AAAI, 2019.

The Proposed IMDer

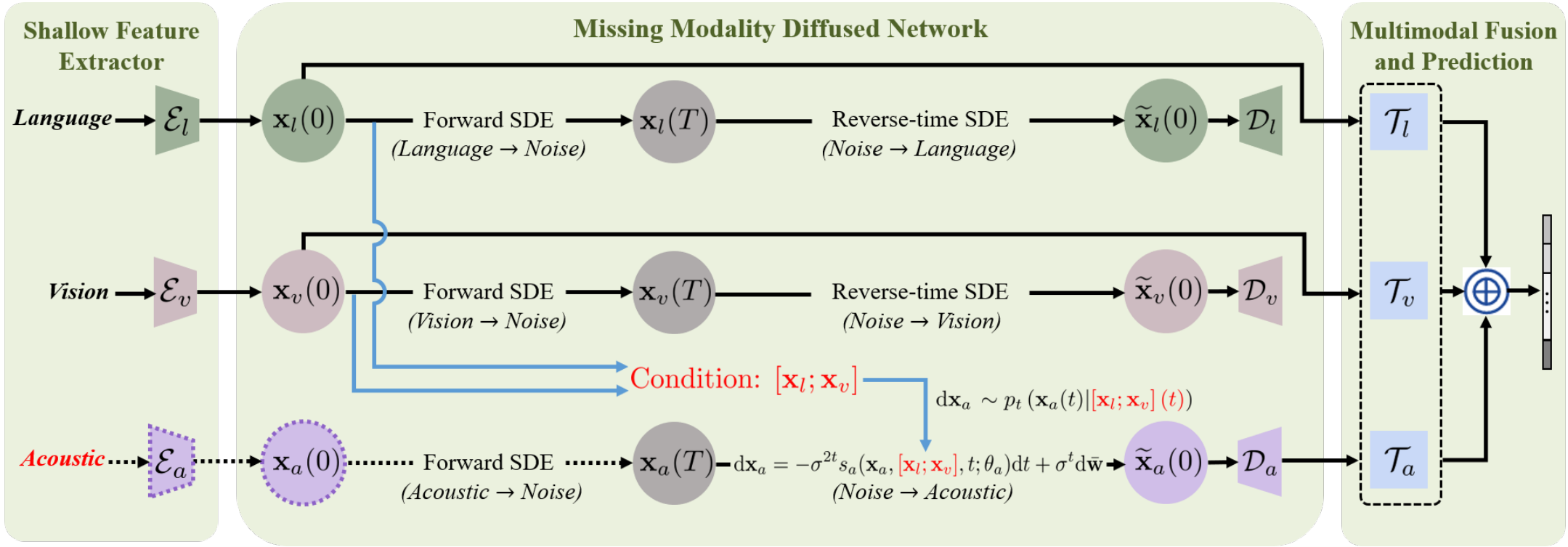
- We propose the Incomplete Multimodality-Diffused emotion recognition (IMDer) method that **maps input random noise to the distribution space of missing modalities** and recovers missing data in accordance with their original distributions.



Our proposed Incomplete Multimodality-Diffused Emotion Recognition (IMDer)

The framework of IMDer

- IMDer maps input random noise to the distribution space of missing modalities and recovers missing data in accordance with their original distributions.
- IMDer utilize the available modalities as prior conditions to guide and refine the recovering process.



The detailed framework of IMDer

Experiments

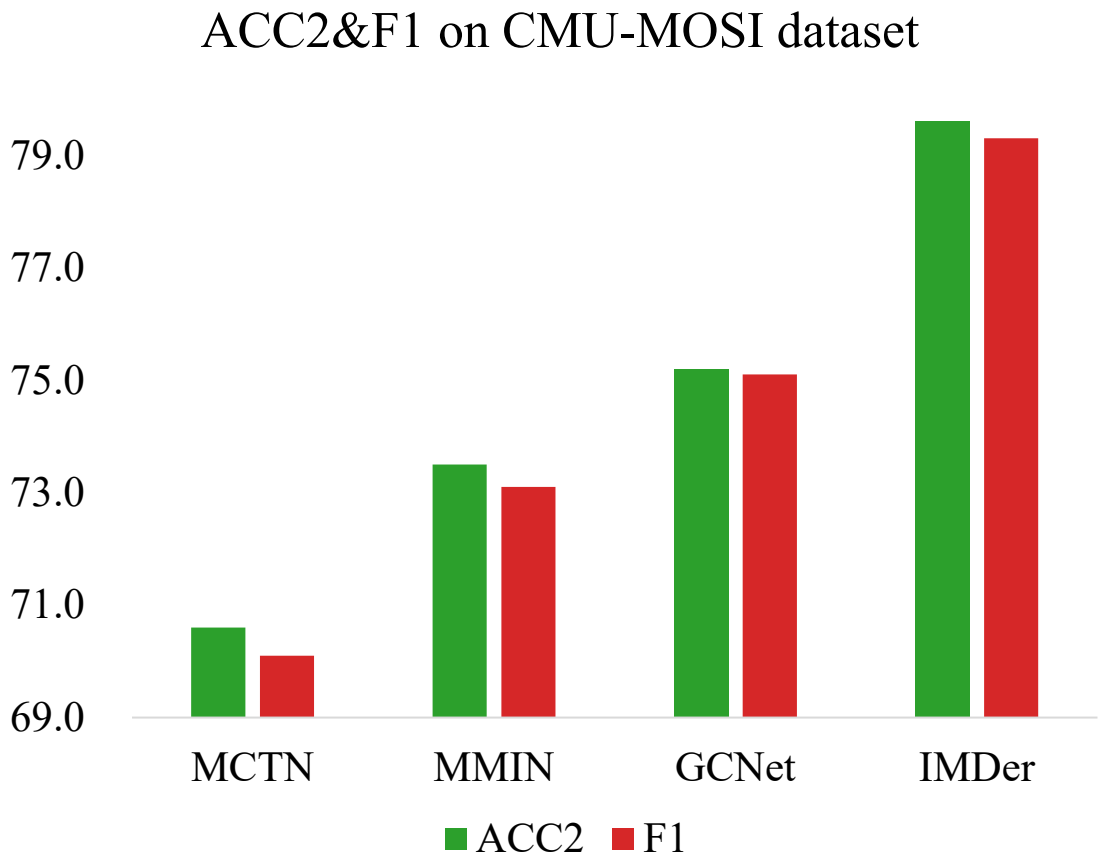
- Datasets
  - **CMU-MOSI**[a] is a MER dataset consisting of 2,199 short monologue video clips (each lasting the duration of a sentence).
  - **CMU-MOSEI**[b] is a large MER dataset, which contains more than 22,000 sentence utterance videos from more than 1000 online YouTube speakers**.**

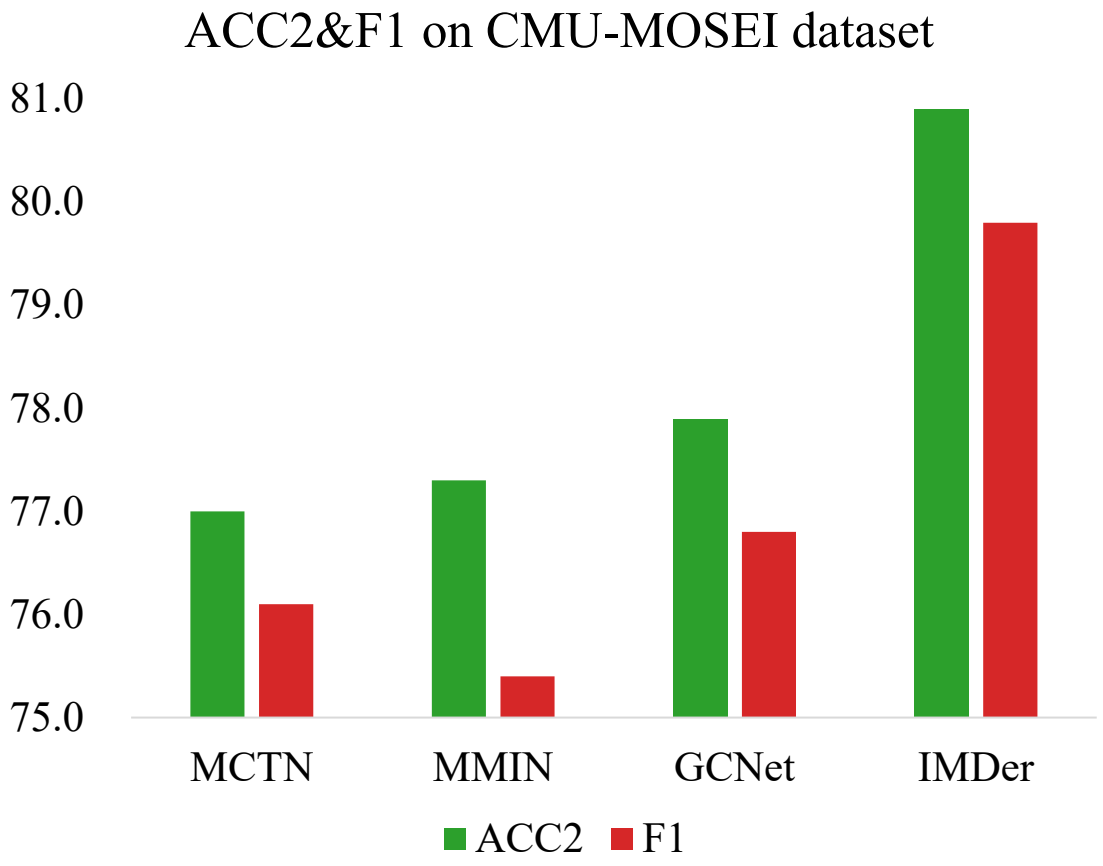
Example face illustration in CMU-MOSEI dataset.

[a] Zadeh, Amir, et al. "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages." IEEE Intelligent Systems. 2016.
[b] Zadeh, Amir, et al. "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph." ACL. 2018.

Experiments

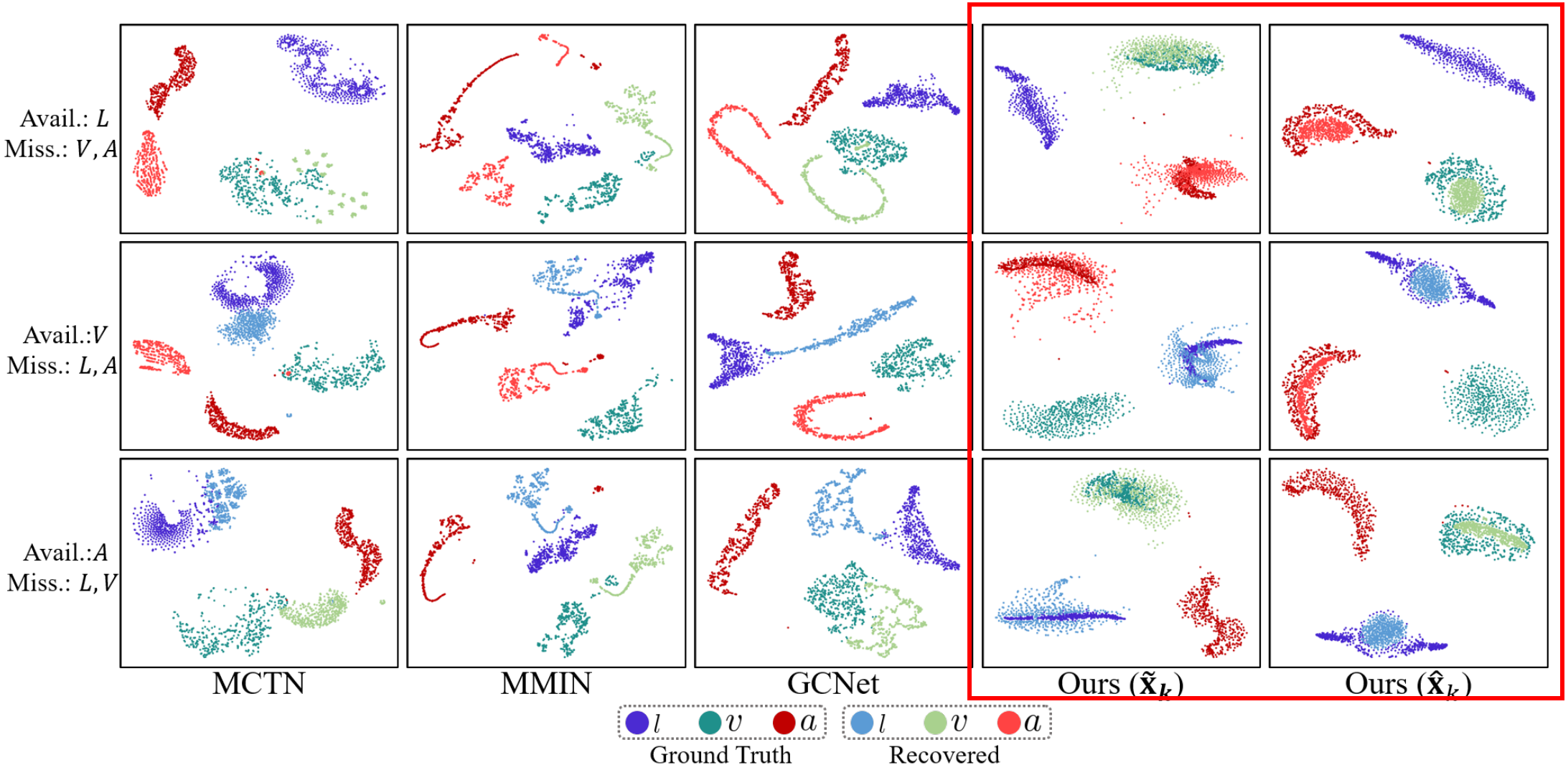
- Quantitative Comparison





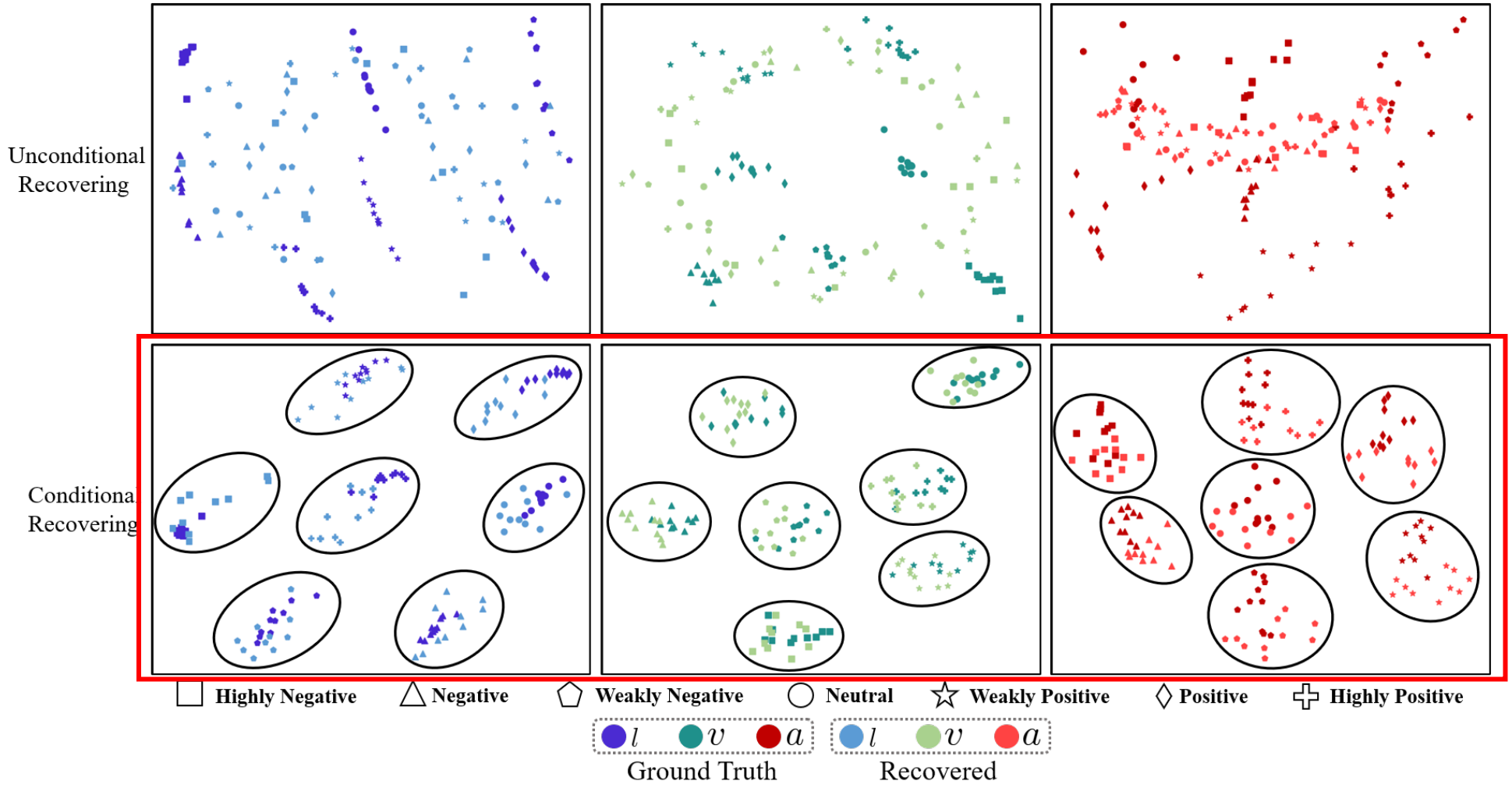
Experiments

- Visualization of recovered modalities



Experiments

- Visualization of the unconditional and conditional modality recovering



Conclusion

- We have proposed a Incomplete Multimodality-Diffused Emotion Recognition (IMDer) for MER under incomplete multimodalities.
- IMDer maps input random noise to the distribution space of missing modalities and recovers missing data in accordance with their original distributions.
- IMDer utilize the available modalities as prior conditions to guide and refine the recovering process.

Thanks for your attention!

Codes:

https://github.com/mdswyz/IMDer