JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Paper ID: TUE-PM-239

Decoupled Multimodal Distilling for Emotion Recognition

Yong Li, Yuanzhi Wang, Zhen Cui
Nanjing University of Science and Technology, Nanjing, China.
{yong.li, yuanzhiwang, zhen.cui}@njust.edu.cn

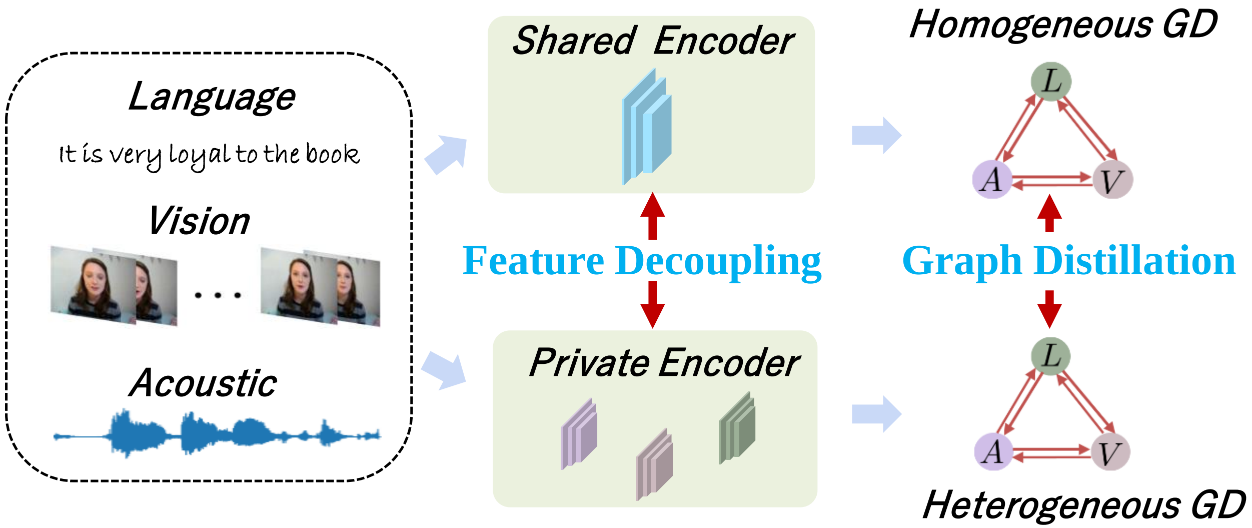


Fig 1. Decoupled and Graph-empowered knowledge distillation for multimodal emotion recognition.

1. Background

- Multimodal Emotion Recognition (MER)

Multimodal emotion recognition (MER) aims to perceive the emotion of humans from video clips.

Video clips involve multimodal temporal data, e.g., natural language, visual actions and acoustic behaviors.

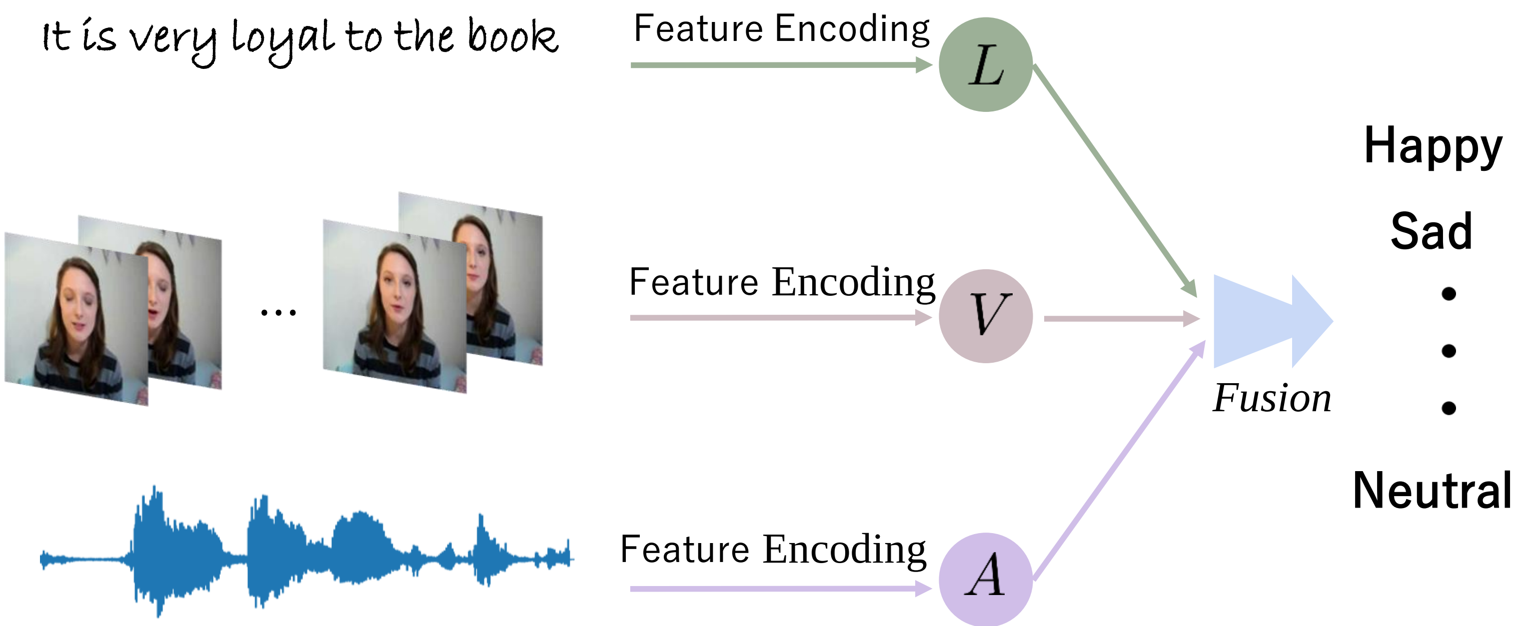
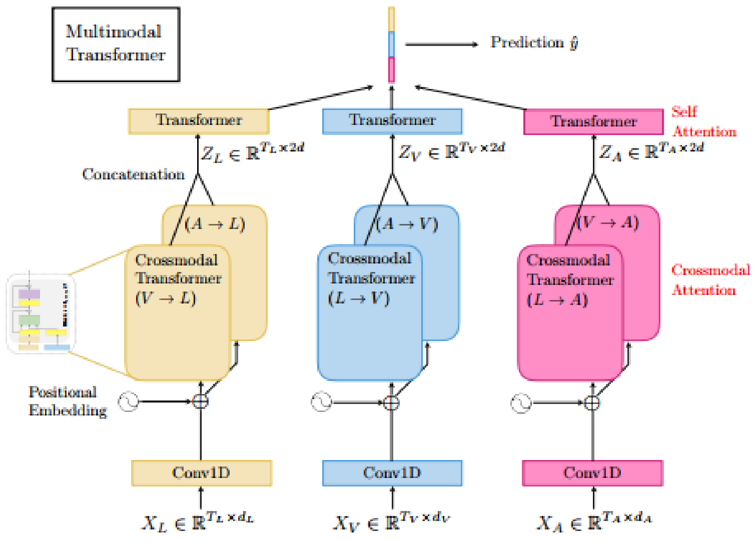


Fig 2. Typical MER pipeline.

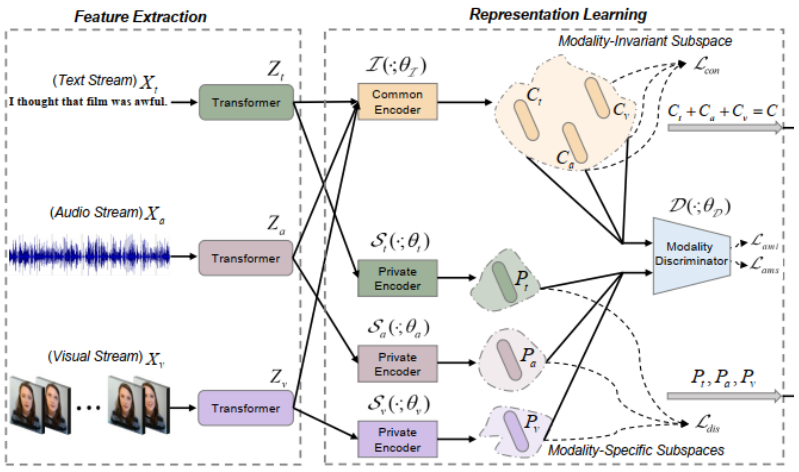
1. Background

- Typical previous methods for MER



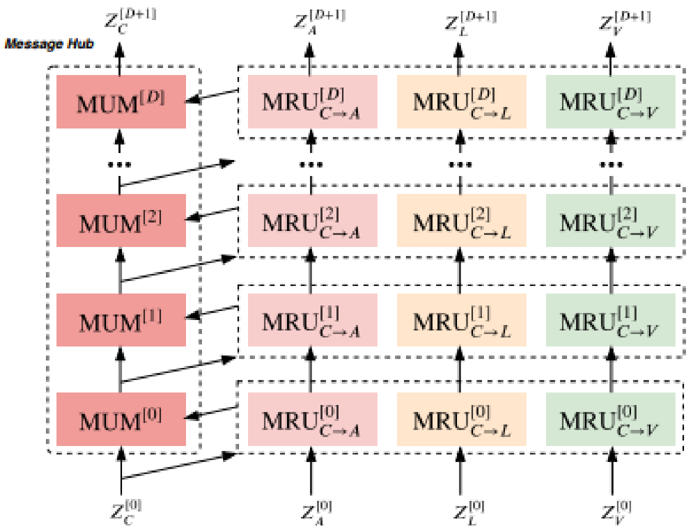
Multimodal Transformer

ACL, 2019 [1]



Feature-Disentangled MER

Multimedia, 2020 [2]



Progressive Modality Reinforcement

CVPR, 2021 [3]

2. Motivation

- Towards small unimodal performance discrepancies

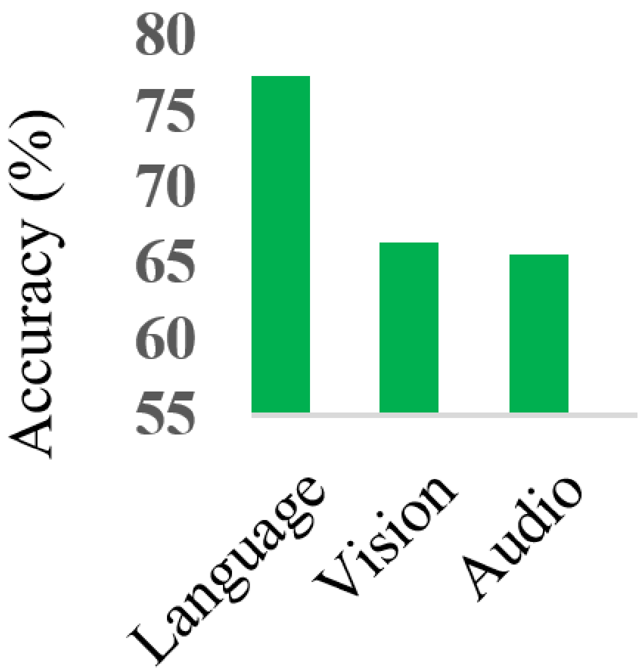
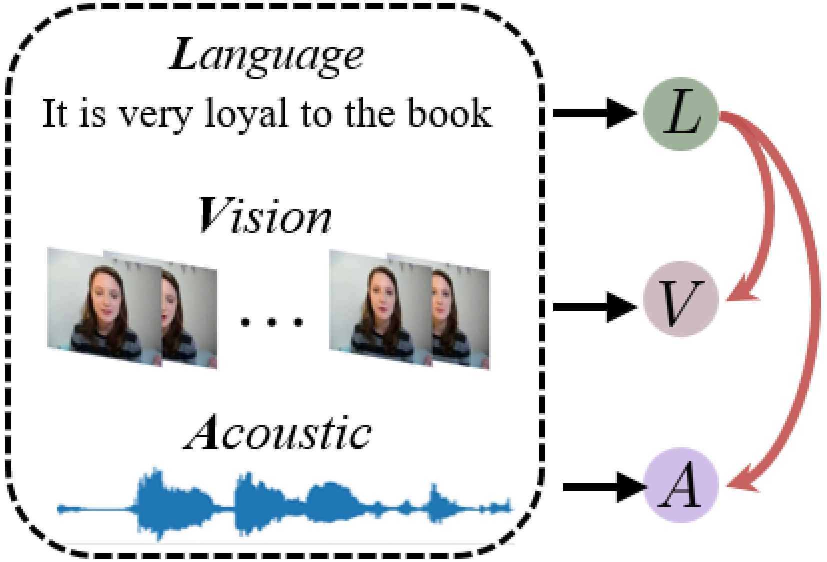


Fig 3. Unimodal accuracy comparison.

- The inherent **multimodal heterogeneities** exist
- The contribution of different modalities varies significantly
  - **Language** excels as it can benefit from a pre-trained model, e.g., BERT
  - **Language** is descriptive, sparse, intrinsically semantic
  - **Vision**/image is redundant
  - **Audio** is quite weak with few semantics

2. Motivation

- Towards small unimodal performance discrepancies



Conventional cross-modal distillation mechanism has **drawbacks**:

- Distillation **direction** **or** **weights** are **cumbersome**
- Multimodal feature **distribution mismatch** hinders the distillation effects

Fig 4. Conventional knowledge distillation for MER.

2. Motivation

- Towards small unimodal performance discrepancies

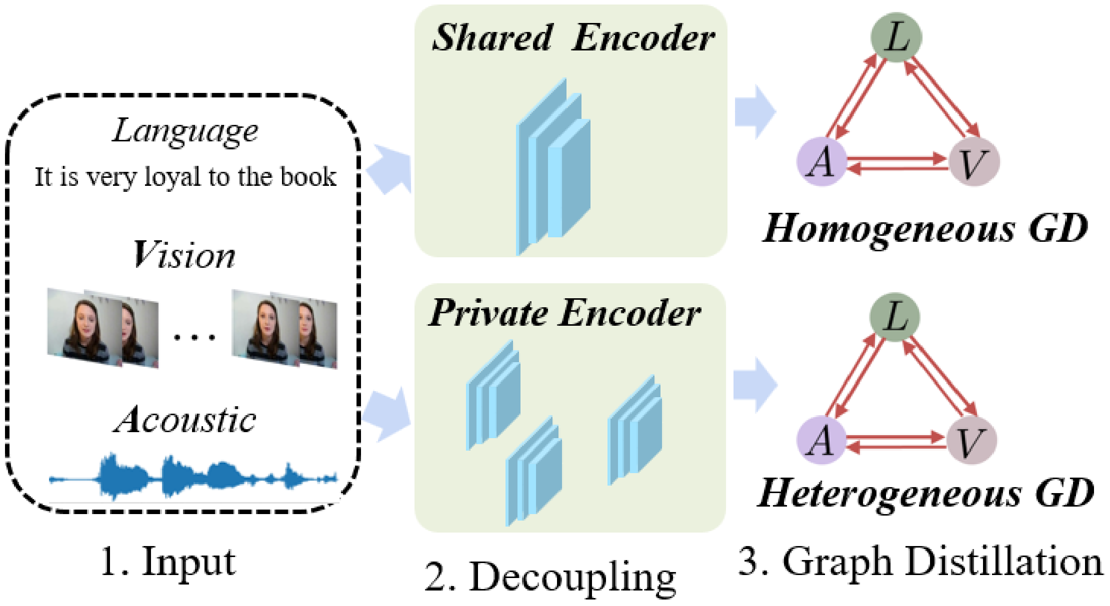


Fig 5. Our proposed Decoupled Multimodal Distillation.

Decoupled multimodal distillation mechanism has benefits:

- Distillation direction and weights can be adaptively learned
- Multimodal heterogeneity can be mitigated via feature decoupling

3. Decoupled Multimodal Distillation

- Feature Decoupling

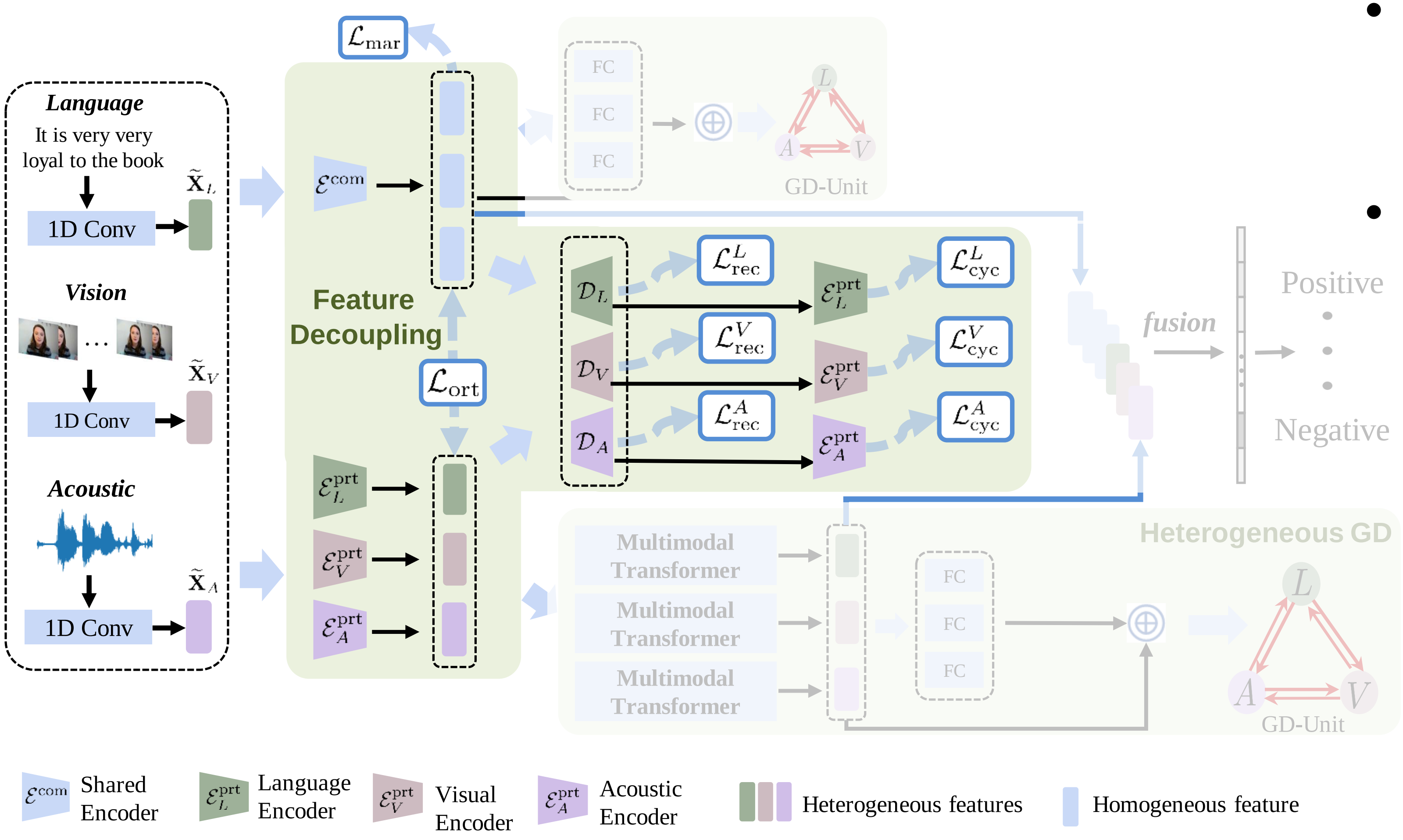


Fig 6. Multimodal Feature Decoupling with self-supervision and contrastive learning.

- Two-level Self-supervised Constraints.
- Margin-based contrastive loss.

Decompose multimodal feature into homo-/ hetero-geneous spaces.

3. Decoupled Multimodal Distillation

- Graph-empowered KD in Homo- space

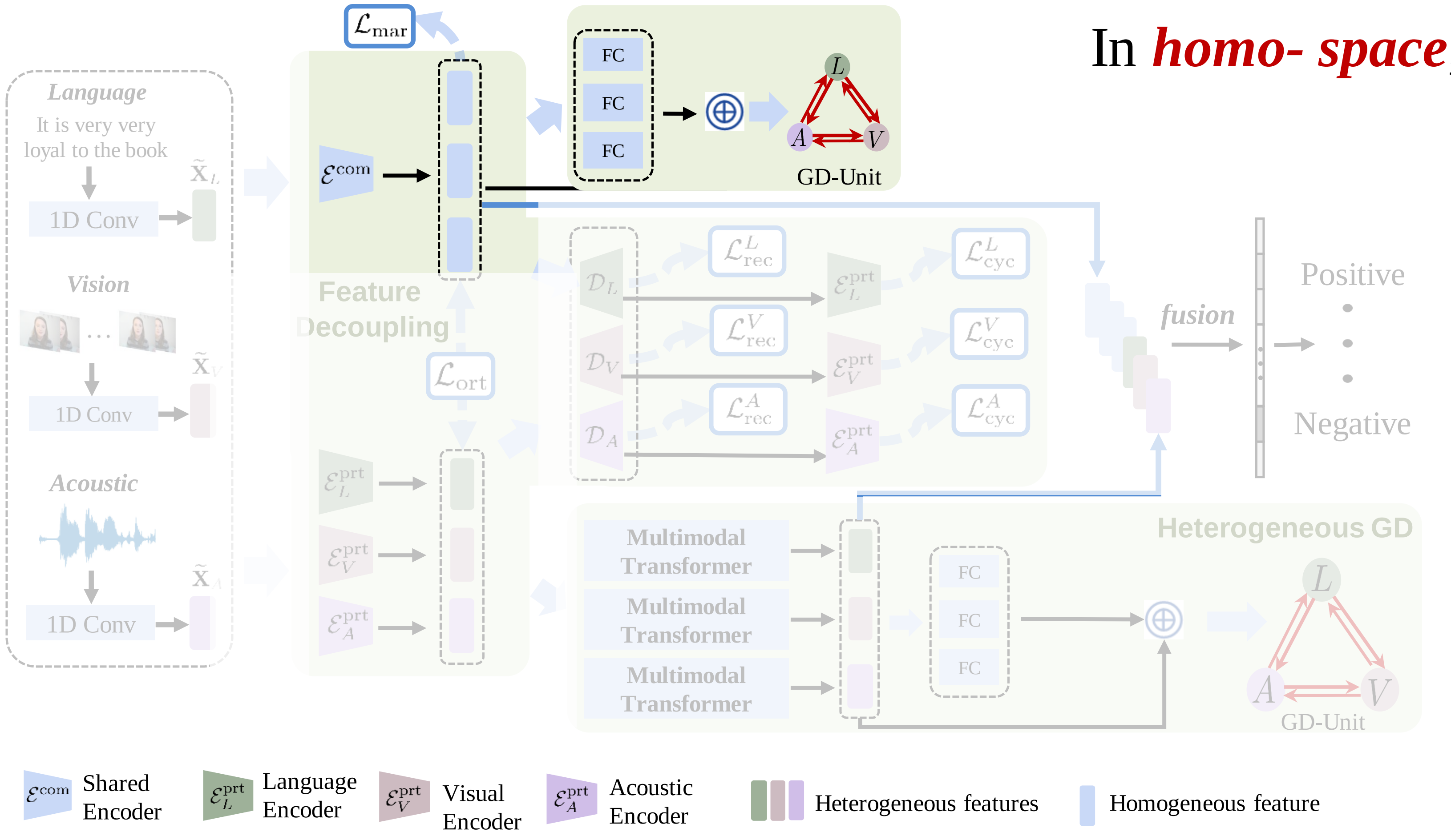


Fig 7. Homogeneous Knowledge Distillation with a Graph Distillation Unit.

In ***homo- space***, KD can be conduct directly.

Graph Distillation:

- Graph ***node***: multimodal feature.
- Graph ***edge***: distillation direction and weight.

3. Decoupled Multimodal Distillation

- Graph-empowered KD in Hetero- space

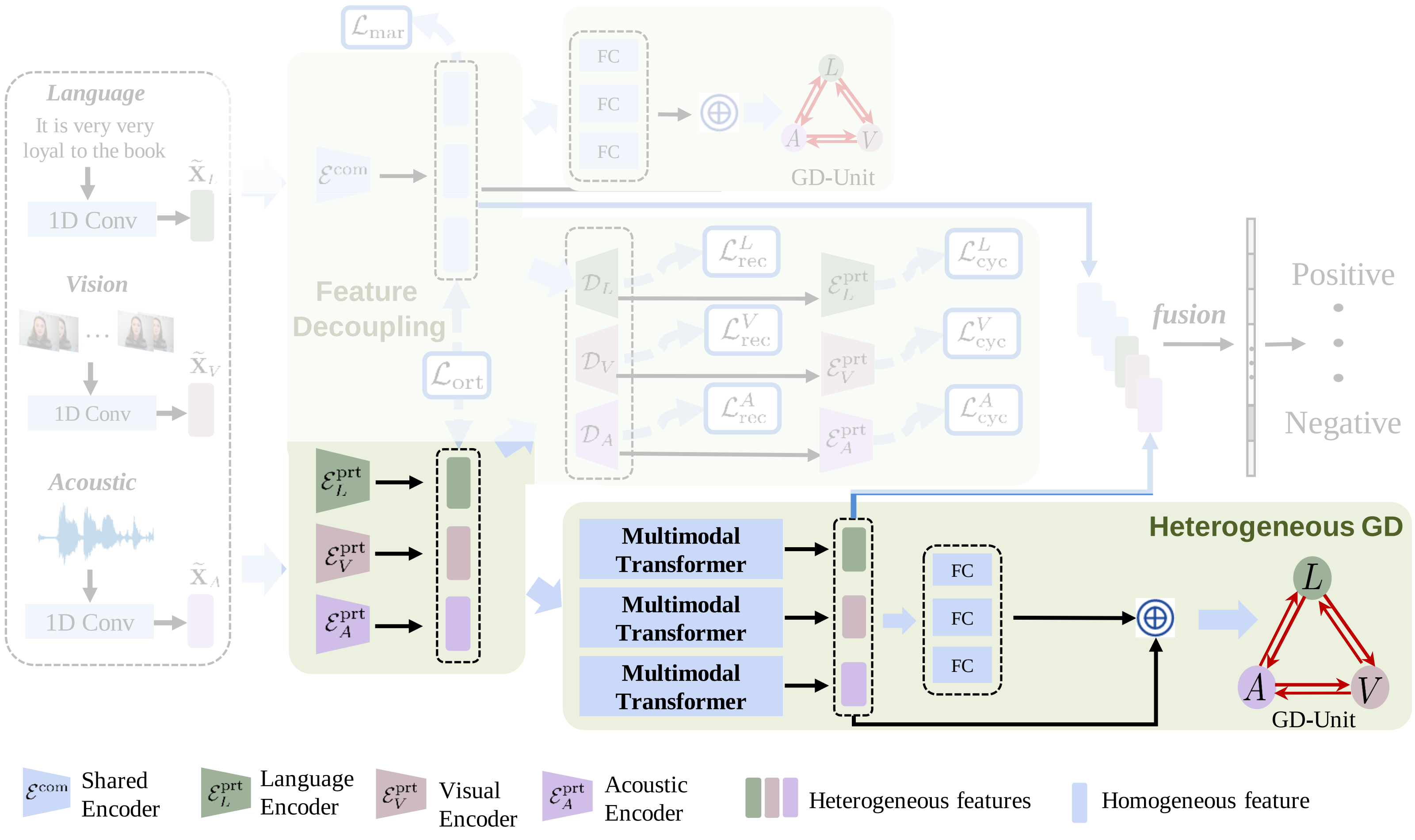


Fig 8. Heterogeneous Knowledge Distillation with a GD-Unit.

In ***hetero- space***, KD should be performed after multimodal feature adaptation.

3. Decoupled Multimodal Distillation

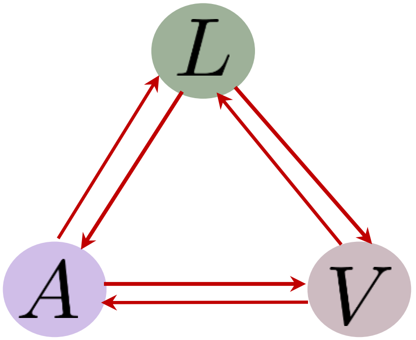
- Graph-empowered KD

Notations:

- A GD-Unit consists of a directed **graph** $\mathcal{G}$
- **Node** v_i denotes a modality
- $w_{i\rightarrow j}$ indicates distillation strength from **i** to **j**
- $\epsilon_{i\rightarrow j}$ denotes distillation loss.

For a target modality, the weighted distillation loss is:

$$\zeta_{:j} = \sum_{v_i \in \mathcal{N}(v_j)} w_{i\rightarrow j} \times \epsilon_{i\rightarrow j}$$



Learnable Graph Edge:

The graph **edge** $w_{i\rightarrow j}$ means distillation strength. We encode the modality logits and the features into the graph edges:

$$w_{i\rightarrow j} = g([[f(\mathbf{X}_i, \theta_1), \mathbf{X}_i], [f(\mathbf{X}_j, \theta_1), \mathbf{X}_j]], \theta_2)$$

Benefits of Graph-empowered KD:

- **Learnable** KD strength
- **Adjustable** KD direction

4. Experiments

- Datasets

  - **CMU-MOSI**[4] is a MER dataset consisting of 2,199 short monologue video clips (each lasting the duration of a sentence).
  - **CMU-MOSEI**[5] is a larger MER dataset, which contains more than 23,500 sentence utterance videos from more than 1000 online YouTube speakers**.**

Fig 9. Example face illustration in CMU-MOSEI dataset.

4. Experiments

- Numeric comparisons

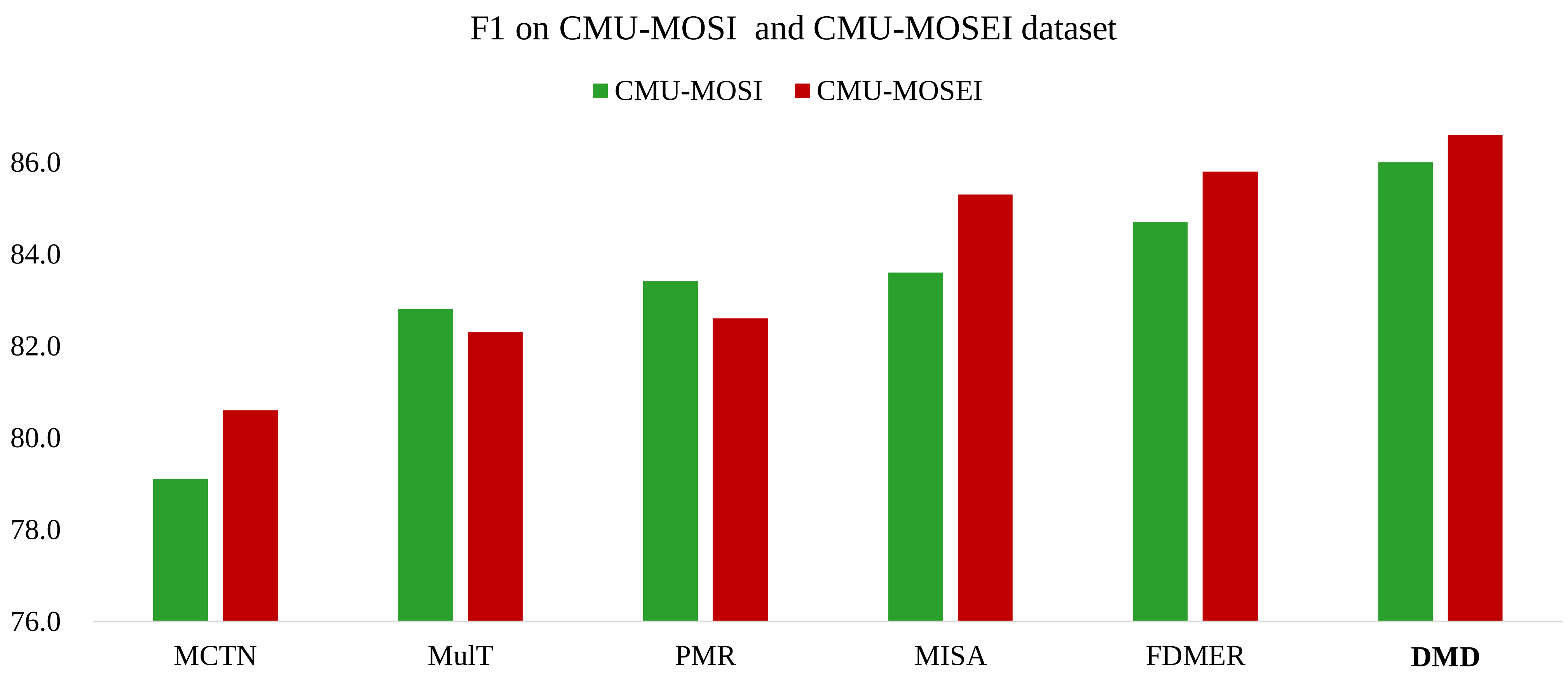


Fig 10. **DMD** consistently obtains **superior** MER accuracy.

4. Experiments

- **Homo**geneous Feature Visualization

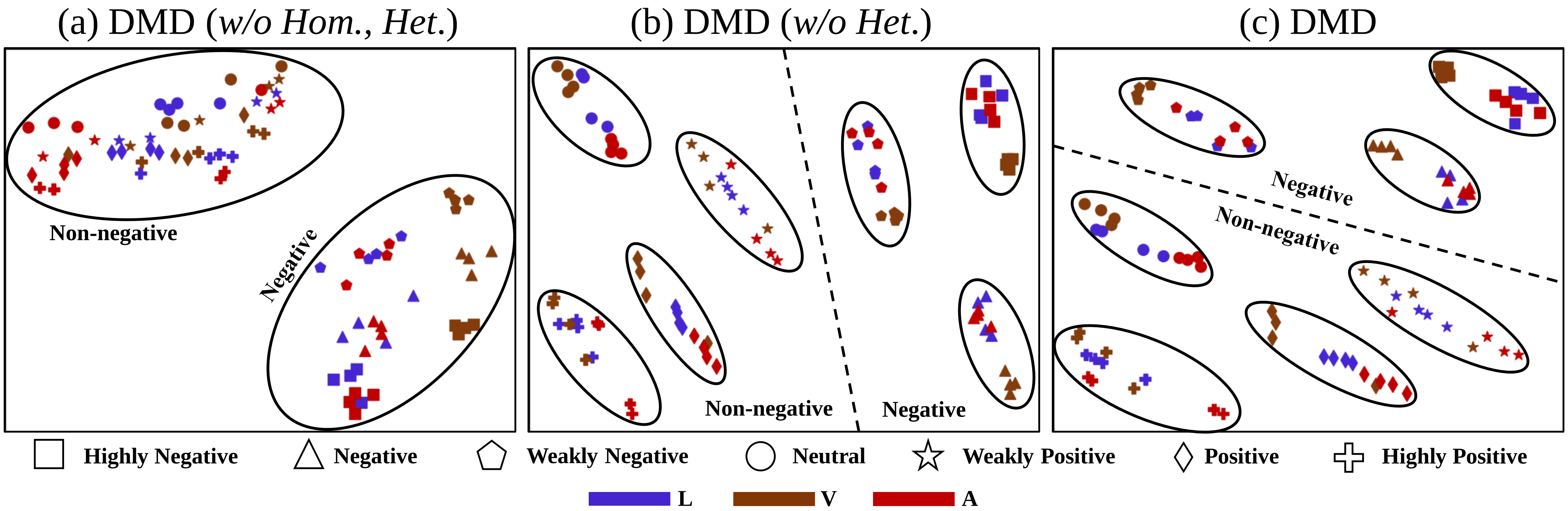


Fig 11. **DMD** shows the promising **emotion category separability** in sub-figure (c).

4. Experiments

- **Hetero**geneous Feature Visualization

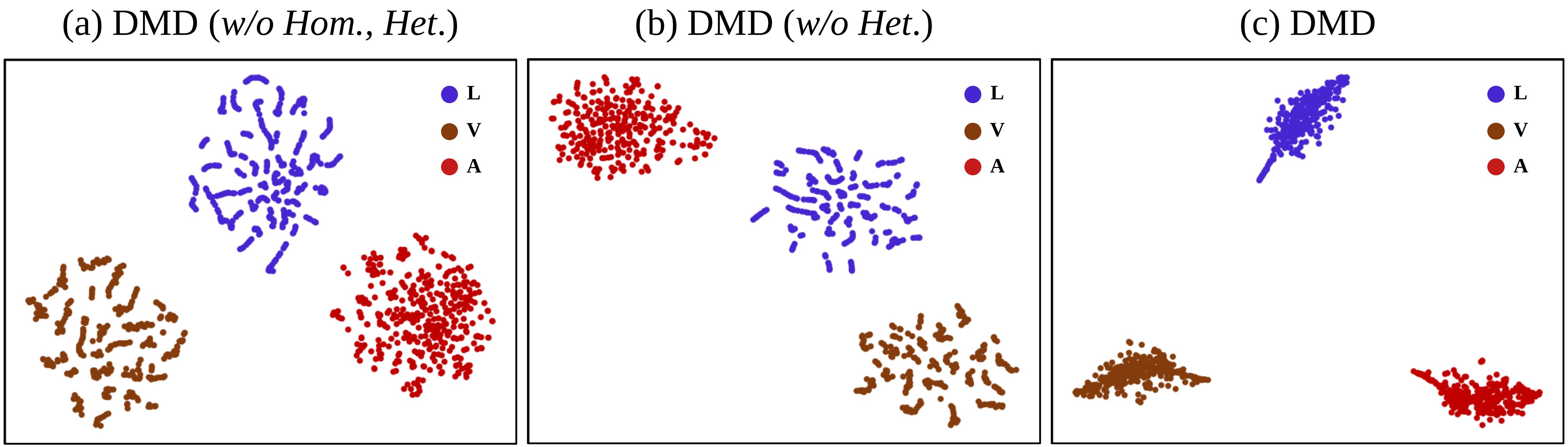


Fig 12. We randomly selected 400 samples for t-SNE visualization.

DMD shows the best **modality separability** in sub-figure (c).

4. Experiments

- Graph **Edge** Visualization

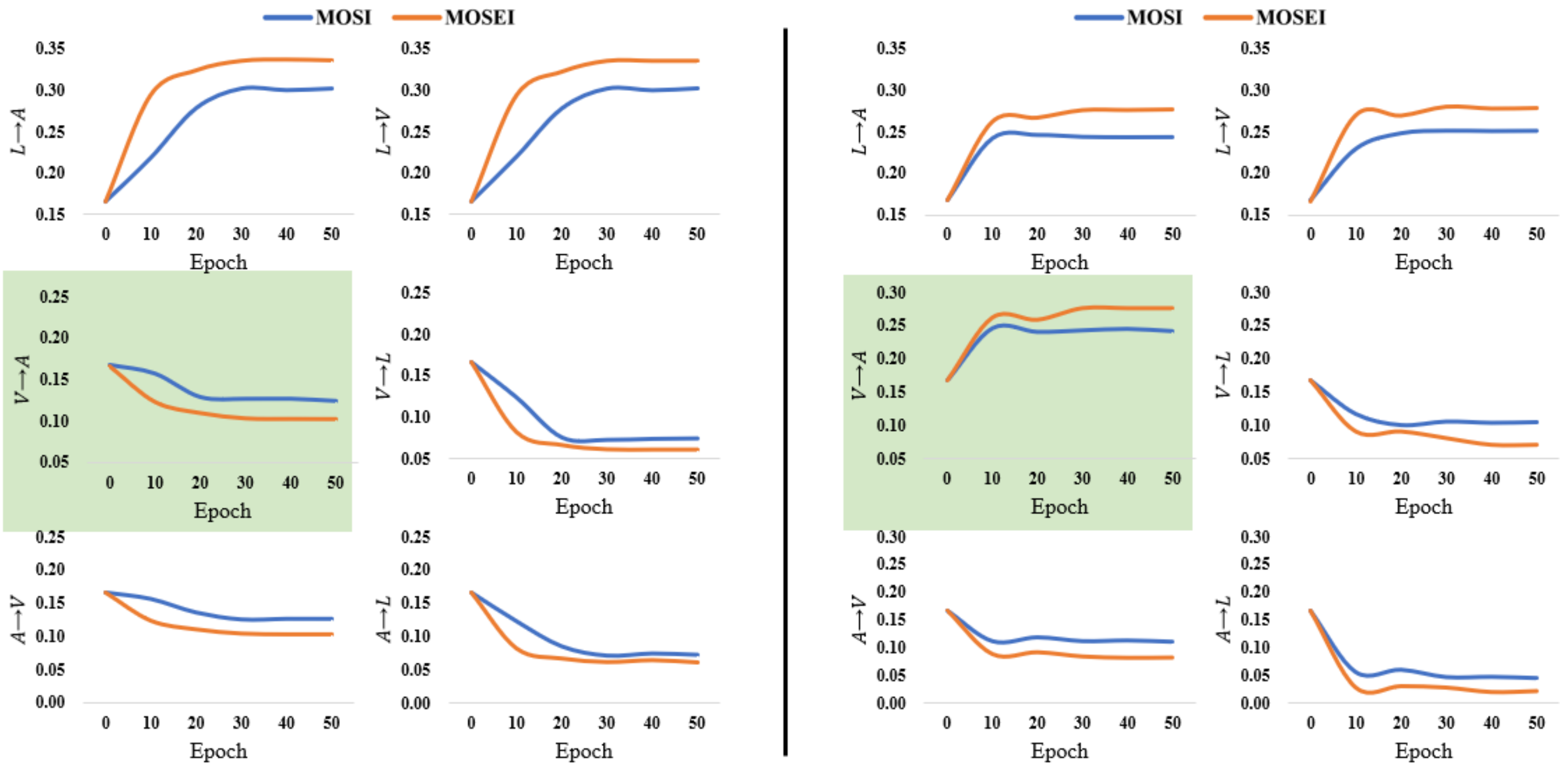
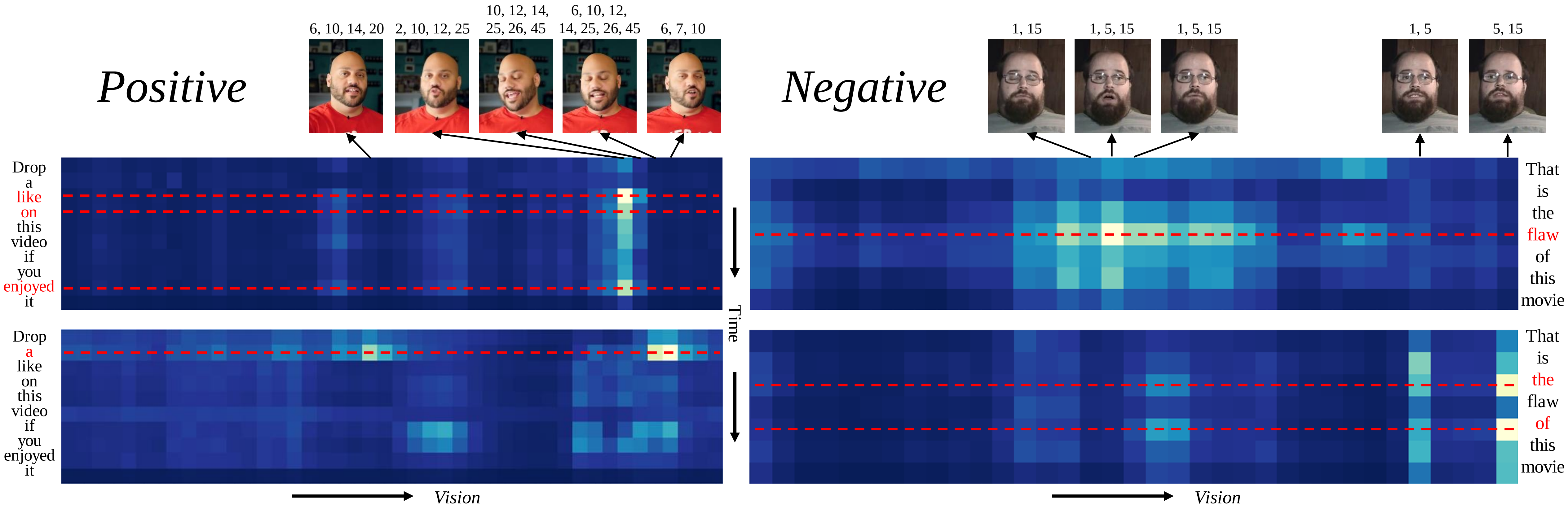


Fig 13. Six graph edge visualization for each MER dataset.

- In the two decoupled spaces, $L \to A$ and $L \to V$ dominates because language contributes most.
- In HeteroGD, $V \to A$ emerges because vision is enhanced a lot via the multimodal transformer mechanism.

4. Experiments

- **Attention** Visualization



(a) Top: **DMD** (**Ours**). Bottom: MulT

(b) Top: **DMD** (**Ours**). Bottom: MulT

Fig 14. In the **top** row, DMD builds reliable correlations between elements across modalities.

5. Conclusion

- We have proposed a Decoupled Multimodal Distillation (DMD) for MER.
- DMD decouples the multimodalities into ***homo***geneous and ***hetero***geneous spaces.
- DMD exploits **graph-empowered Knowledge Distillation** for robust MER.

Thanks for your attention!

Public Code:

https://github.com/mdswyz/DMD

6. Reference

- [1] Tsai *et al*. Multimodal transformer for unaligned multimodal language sequences. ACL, 2019
- [2] Hazarika *et al*. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. MM. 2020
- [3] Lv *et al*. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. CVPR, 2021
- [4] Zadeh *et al*. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. IEEE Intelligent Systems, 2016
- [5] Zadeh *et al*. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. ACL, 2018