Text-Guided Nonverbal Enhancement based on Modality-Invariant and -Specific Representations for Video Speaking Style Recognition

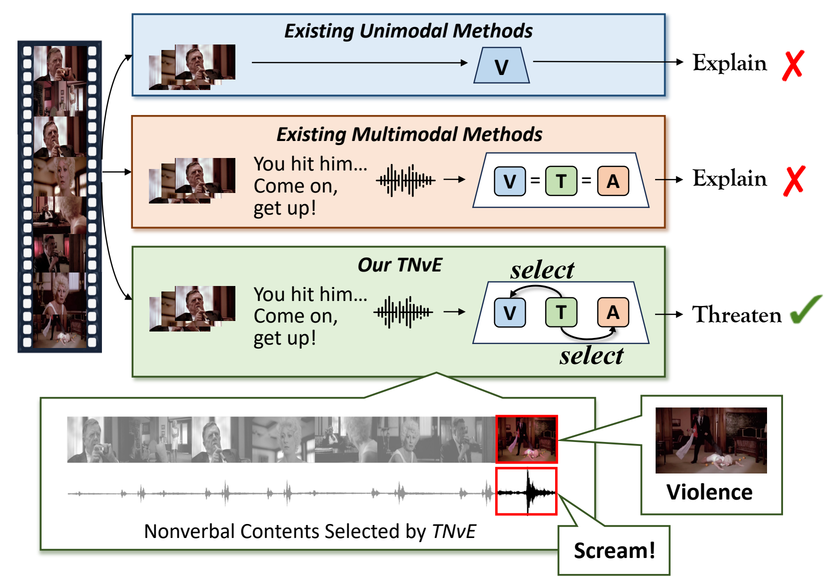
AAAI 2025

Beibei Zhang, Tongwei Ren, Gangshan Wu

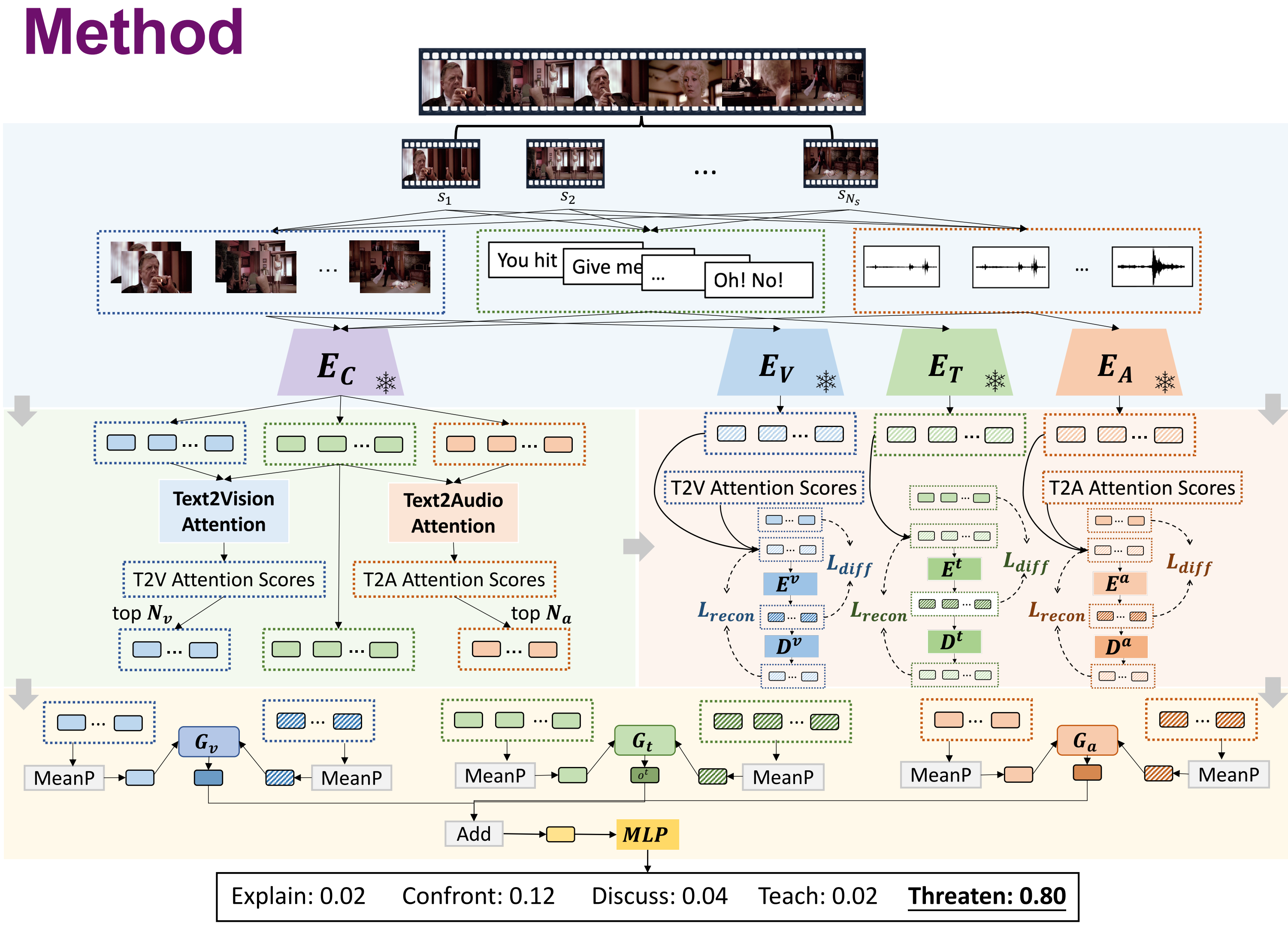
State Key Laboratory for Novel Software Technology, Nanjing University

Introduction

Video speaking style recognition (VSSR) aims to classify different types of conversations in videos, which is a fine-grained video understanding task. We propose a text-guided nonverbal enhancement method, **TNvE**, which is composed of a text-guided nonverbal representation selection module and a modality-invariant and -specific representation decoupling module, significantly improving the performance of VSSR and achieves a new state-of-the-art.



Method



There are four main steps in TNvE: 1) Firstly the input video is segmented into multiple shots, from which modality-invariant and -specific multimodal representations are extracted; 2) A limited number of critical nonverbal representations are selected with the guide of text in the modality-invariant embedding space. And invariant and specific representations of selected shots are preserved; 3) After that, a representation decoupling module is applied to minimize redundancy between modality invariant and -specific representations; and 4) Finally invariant and specific representations of the same modality are adaptively fused and all multimodal representations are then aggregated to predict the speaking style.

Experiments

Dataset: LVU-VSSR, LVU-VSRR

Metrics: Accuracy, F1-score, Precision and Recall

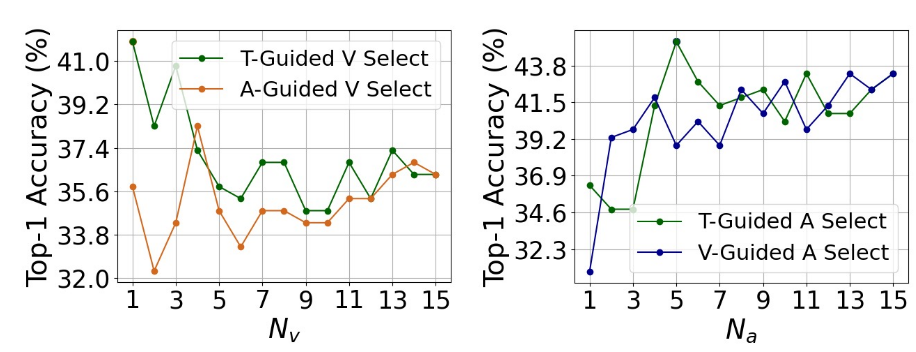
Comparison with the SOTA: TNvE is superior to all VSSR methods in all metrics.

Method	Acc	F1	P	R	WF1	WP
Unimodal						
ObjTrans(Wu and Krahenbuhl 2021)	40.3	35.7	36.2	36.4	39.1	38.7
ViS4mer (Islam and Bertasius 2022)	38.3	32.9	35.3	34.3	36.3	37.2
S5 (Wang et al. 2023b)	42.1	-	-	-	-	-
Multimodal						
TFN (Zadeh et al. 2017)	30.8	20.1	18.6	24.1	25.8	24.1
MulT (Tsai et al. 2019)	46.8	40.2	43.0	42.7	43.7	44.8
Bert-MAG (Rahman et al. 2020a)	44.8	39.9	45.8	42.8	40.8	46.4
LF-VILA (Sun et al. 2022)	40.3	31.9	31.1	34.1	37.6	36.6
DMD (Li, Wang, and Cui 2023)	40.3	26.7	25.1	32.5	34.0	32.8
Movie2Scenes (Chen et al. 2023)	42.2	-	-	-	-	-
LMP (Argaw et al. 2023)	44.4	-	-	-	-	-
MMSF (Zhang et al. 2023)	50.2	45.0	48.0	44.5	49.1	49.5
MA-LLM (He et al. 2024)	41.2	36.4	40.4	38.1	39.0	42.4
LSSD (Singh et al. 2024)	50.8	-	-	-	-	-
TNvE (Ours)	**56.7**	**51.7**	**56.8**	**53.3**	**54.8**	**57.9**

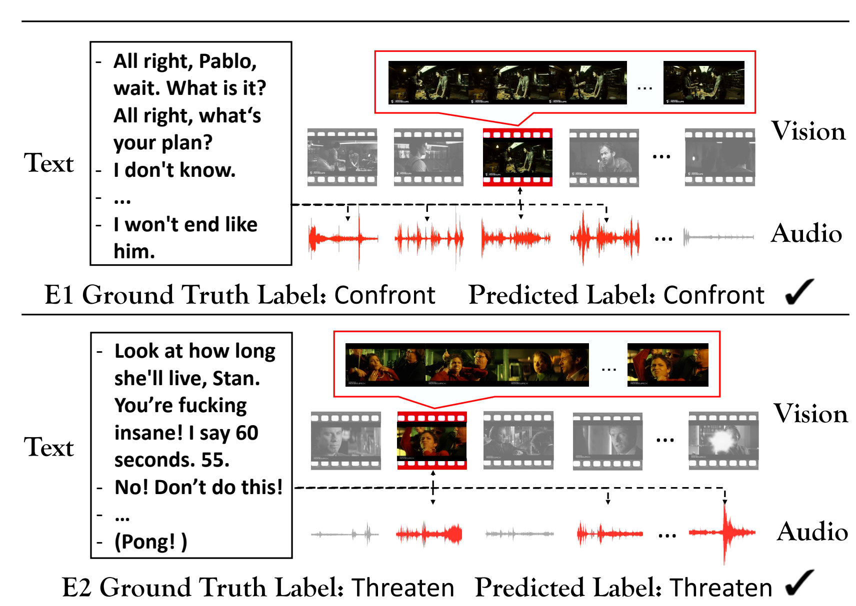
Ablation Study: We conduct multiple ablation experiments. The results demonstrate that text-guided selection can boost VSSR performance and representation decoupling is necessary for comprehensive multimodal understanding.

v	a	t	w/o TNvRS			w/ TNvRS		
			Acc	R	WF1	Acc	R	WF1
✓			36.3	30.9	34.2	41.8	34.7	38.4
	✓		43.3	39.7	43.0	45.3	39.1	43.1
		✓	50.3	47.4	48.6	50.3	47.4	48.6
✓	✓		41.3	37.8	40.7	37.8	31.4	34.9
✓		✓	48.3	45.5	47.3	51.2	49.2	49.7
	✓	✓	50.3	47.0	50.2	50.8	49.6	49.1
✓	✓	✓	47.3	45.5	46.8	54.2	53.4	52.8

M-I	M-S	Diff	Recon	Fusion	Acc	F1	WF1	WP
✓				-	54.2	51.4	52.8	54.3
	✓			-	50.3	41.9	46.7	54.7
✓	✓			Add	53.2	47.5	52.0	53.3
✓	✓	✓		Add	54.2	49.7	52.9	57.3
✓	✓	✓	✓	Add	55.7	52.0	54.5	55.0
✓	✓	✓	✓	Concat	52.2	47.7	51.4	53.2
✓	✓	✓	✓	Gate	56.7	51.7	54.8	57.9



Qualitative Analysis: TNvE can effectively leverage text to select critical nonverbal cues to enhance the recognition accuracy of VSSR.



zhangbb@smail.nju.edu.cn
rentw@nju.edu.cn

NANJING UNIVERSITY

MAGUS MediA recoGnition and UnderStanding