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

Beibei Zhang, Tongwei Ren*, Gangshan Wu

State Key Laboratory for Novel Software Technology, Nanjing University
zhangbb@smail.nju.edu.cn, {rentw, gswu}@nju.edu.cn

Abstract

Video speaking style recognition (VSSR) aims to classify different types of conversations in videos, contributing significantly to understanding human interactions. A significant challenge in VSSR is the inherent similarity among conversation videos, which makes it difficult to distinguish between different speaking styles. Existing VSSR methods commit to providing available multimodal information to enhance the differentiation of conversation videos. Nevertheless, treating each modality equally leads to a suboptimal result for these methods due to text is inherently more aligned with conversation understanding compared to nonverbal modalities. To address this issue, we propose a text-guided nonverbal enhancement method, TNvE, which is composed of two core modules: 1) a text-guided nonverbal representation selection module employs cross-modal attention based on modality-invariant representations, picking out critical nonverbal information via textual guide; and 2) a modality-invariant and -specific representation decoupling module incorporates modality-specific representations and decouples them from modality-invariant representations, enabling a more comprehensive understanding of multimodal data. The former module encourages multimodal representations close to each other, while the latter module provides unique characteristics of each modality as a supplement. Extensive experiments are conducted on long-form video understanding datasets to demonstrate that TNvE is highly effective for VSSR, achieving a new state-of-the-art.

Introduction

Video speaking style recognition aims to recognize different categories of human conversations in videos, offering a critical perspective to understand human interactions. It plays a vital role in many applications, such as video content understanding (Wu and Krahenbuhl 2021), conversation head generation (Zhou et al. 2022) and virtual agent design (Aneja et al. 2021). As a result, VSSR has attracted considerable attention in recent years.

As shown in Figure 1, existing VSSR methods can be broadly categorized into two main types: unimodal and multimodal approaches. Unimodal methods (Wu and Krahenbuhl 2021; Islam and Bertasius 2022; Fish, Weinbren, and Gilbert 2022; Wang et al. 2023b) focus on developing sophisticated visual networks to capture robust spatio-temporal relationships within dynamic video contents to reveal the speaking styles. However, the visual contents of conversation videos, which are composed of consistent face-to-face talking shots, tend to be too similar to distinguish between different speaking styles, leading to a performance decline in VSSR (Zhang et al. 2023). As a result, multimodal methods (Sun et al. 2022; Chen et al. 2023; Zhang et al. 2023; Argaw et al. 2023; He et al. 2024) commit to exploring multimodal available information for VSSR as a complement to visual representation, aiming to enhance the differentiation of conversation videos.

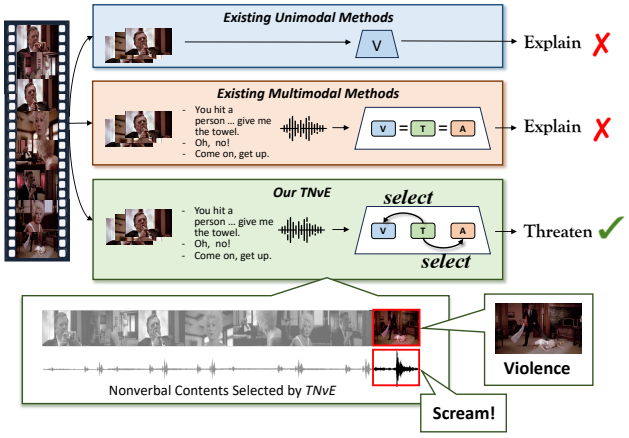

Figure 1: Comparison between existing unimodal methods, multimodal methods and our TNvE for VSSR. Here, V, A and T refers to visual, acoustic and textual modalities, nonverbal contents in red boxes are critical ones selected by TNvE.

Despite the progress made by multimodal methods, they treat each modality equally, neglecting the fact that different modalities contribute to VSSR to varying extents. Text is inherently superior to nonverbal modalities for VSSR due to textual dialogue contains dense semantic information highly related to the conversation. In contrast, nonverbal contents, typically involving face-to-face talking shots and moderate volume and pitches, tend to remain consistent

*Corresponding author.

for most of the time, with few key behaviors that are indicative of speaking styles interspersing throughout. The similarity of nonverbal contents poses a notable challenge in discriminating different speaking styles, in which context treating text and nonverbal modalities equally will bring in negative noise rather than positive complementary during multimodal fusion.

Motivated by these observations, we propose a text-guided nonverbal enhancement method, TNvE, to address the aforementioned issue. As shown in Figure 1, the core concept of TNvE is applying text as the guide to identify and select critical nonverbal information, *e.g.,* physical violence and scream related to the speaking style of "Threaten", thereby reducing redundancy and noise in nonverbal data that leads to the false prediction "Explain". Specifically, we compute cross-modal attention between textual and nonverbal representations, considering nonverbal cues which are close to texts as valuable information for VSSR. Due to cross-modal matching suffers from multimodal distribution gap (Liang et al. 2021), we extract modality-invariant multimodal representations that distribute in a common embedding space to perform cross-modal attention more effectively, with the help of a pre-trained multimodal binding model (Girdhar et al. 2023).

Since both text-guided nonverbal representation selection and modality-invariant representation extraction inherently encourage multimodal representations close to each other in the distribution space, it potentially leads to a lack of the unique characteristics of each modality. To overcome this issue, the second core component of TNvE is to construct modality-specific multimodal representations, which distribute in private embedding spaces and capture the unique attributes of each modality, as a supplement. The key practice to utilize modality-invariant and -specific representations concurrently is to achieve representation decoupling to avoid information duplication. To this end, after extracting modality-specific representations with unimodal encoders, a disparity constraint is applied on modality-invariant and -specific representations to ensure that they capture information of different aspects. In addition, reconstruction based on the decoupled modality-specific representations is performed to ensure them retaining critical characteristics without losing useful information during decoupling. Finally, invariant and specific representations of the same modality are adaptively fused via a gate, followed by multimodal representation aggregation, to achieve comprehensive multimodal understanding.

The main contributions of our work can be summarized as follows:

- We propose a novel method, TNvE, for VSSR, which enhances nonverbal modalities by leveraging text to select critical nonverbal cues based on modality-invariant representations, distinguishing the contribution of different modalities to VSSR and overcoming the difficulty in discriminating similar conversation videos more effectively.
- We incorporate modality-specific representations and decouple them with modality-invariant representations to capture the unique characteristics of each modality, thereby contributing to comprehensive multimodal understanding.

Related Work

Video Speaking Style Recognition. There are two mainstream VSSR solutions: unimodal and multimodal methods. Unimodal methods contribute to effectively capturing the spatio-temporal dependencies within videos via designing flexible visual networks using transformer structure (Wu and Krahenbuhl 2021) and state-space modules (Islam and Bertasius 2022; Wang et al. 2023b). To deal with the inherent similarity between visual contents, multimodal methods commit to incorporating multimodal valid information as a complement with the help of large-scale pre-training (Sun et al. 2022; Chen et al. 2023; Argaw et al. 2023) and external knowledge (Zhang et al. 2023; He et al. 2024). Existing multimodal methods treat each modality equally, which can hinder the performance of multimodal fusion as different modalities make contributions of different extents. In contrast, our TNvE method is designed to utilize text as guide to identify critical nonverbal contents, thereby enhancing multimodal representations and improving multimodal fusion outcomes.

Text-Centered Multimodal Learning. Text has been demonstrated to surpass nonverbal modalities in various researches (Wang et al. 2019; Sun et al. 2020; Rahman et al. 2020b; Wang et al. 2023a). This superiority can be attributed to two main factors: 1) text inherently contains more semantics that align closely with human cognition; and 2) pre-trained language models have achieved remarkable performance in capturing and utilizing semantic information in text. To capitalize on the advantages of text, existing methods strengthen nonverbal representations by integrating them with valid textual representations (Wang et al. 2019; Sun et al. 2020) or incorporating them into pre-trained language models (Rahman et al. 2020b; Wang et al. 2023a). Different from these methods, TNvE enhances nonverbal representations by picking out critical nonverbal contents guided by text, rather than directly fusing them with text, averting nonverbal redundancy and noise in conversation videos more effectively.

Multimodal Representation Decoupling. The distribution gap in heterogeneous modalities poses as a great challenge for modeling multimodal interactions (Mai, Hu, and Xing 2020; Liang et al. 2021). Decoupling multimodal representations into modality-invariant and -specific spaces is considered as an effective approach to solve this challenge. Modality-invariant representations distribute in a common embedding space that minimizes the modality gap, while modality-specific representations distribute in distinct, private embedding spaces to capture unique characteristics of each modality. Existing approaches (Hazarika, Zimmermann, and Poria 2020; Yang et al. 2022; Li, Wang, and Cui 2023; Yu et al. 2021) train invariant and specific encoders on raw multimodal features to build modality-invariant and -specific embedding spaces from scratch. The proposed

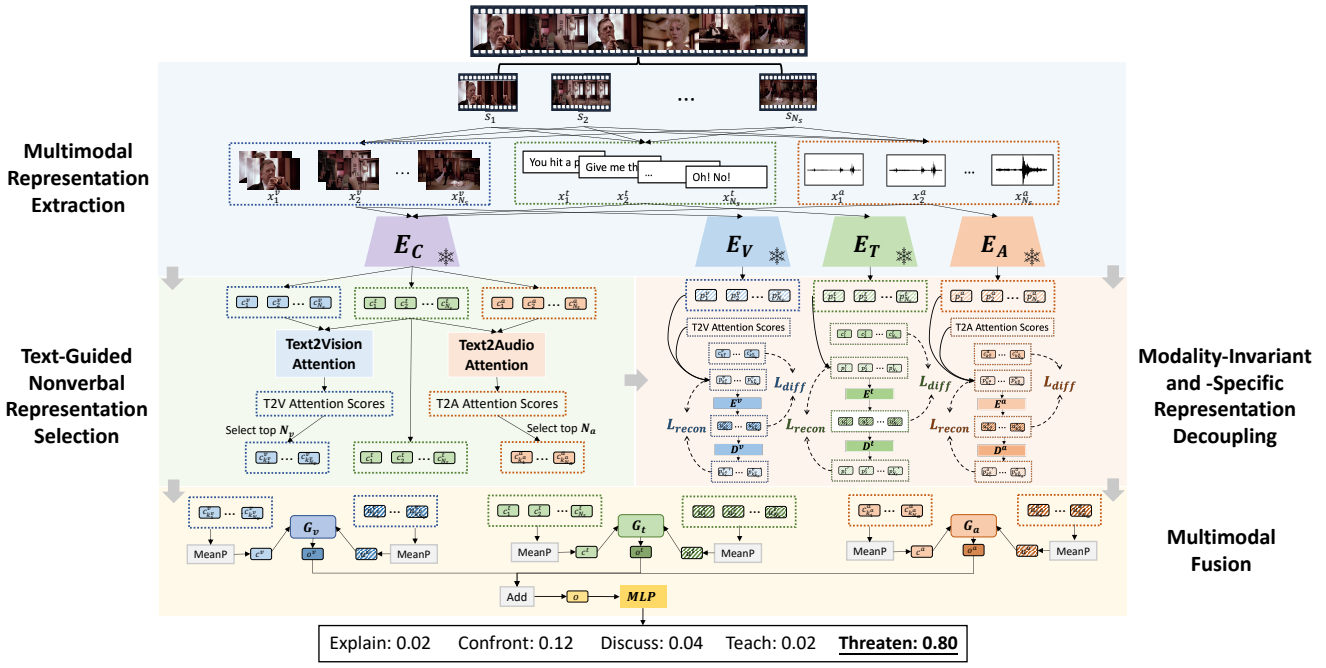

Figure 2: An overview of TNvE. Here, v, a and t refer to visual, acoustic and textual modalities respectively, c, p, u, $p\prime$ and o represent modality-invariant, modality-specific, decoupled modality-specific, reconstructed modality-specific and fused modality-invariant and -specific representations, MeanP represents mean pooling.

TNvE innovatively conduct decoupling based on modality-invariant and -specific representations extracted by advanced multimodal pre-trained models.

Method

The overall framework of TNvE is shown in Figure 2, which outputs the speaking style classification result of the input conversation video. 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. Consequently, modality-invariant and -specific representations of these 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 via a gate and all multimodal representations are then aggregated to predict the speaking style.

Multimodal Representation Extraction

The preposed step for nonverbal representation selection is to set an appropriate selection unit, which is required to satisfy two key criteria: 1) low coupling that each unit differs from the others; and 2) high cohesiveness that each unit contains sufficient meaningful information. Shot, which is known as the fundamental unit of video production (Bordwell, Thompson, and Smith 2010; Argaw et al. 2023), inherently contains distinct contents, suitable to be the selection unit. Thus, the input video V is firstly divided into a sequence of shots $S = \{s_1, s_2, ..., s_{N_s}\}$ using a shot boundary detection model. Video and audio of each shot are regarded as the visual and acoustic raw data $X^v = \{x_1^v, ..., x_{N_s}^v\}$ and $X^a = \{x_1^a, ..., x_{N_s}^a\}$. Since video is the only one input, audio is converted to subtitle using an automatic audio recognition model, yielding textual raw data $X^t = \{x_1^t, ..., x_{N_s}^t\}$.

To address the distribution gap among heterogeneous modalities, modality-invariant representations, which distribute in a common embedding space, are extracted to facilitate the following cross-modal interactions. Benefited from the advance of large-pretrained models, we extract modality-invariant multimodal representations $C^v = \{c_1^v, ..., c_{N_s}^v\} \in R^{N_s \times d^c}$, $C^a = \{c_1^a, ..., c_{N_s}^a\} \in R^{N_s \times d^c}$ and $C^t = \{c_1^t, ..., c_{N_s}^t\} \in R^{N_s \times d^c}$ from X^v, X^a and X^t leveraging a pre-trained multimodal binding model E_C, which aligns multimodal features via contrast learning to obtain a common embedding space involved multiple modalities.

Contrast learning encourages minimizing the distances among multimodal features and preserving consistent semantic information, for which modality-invariant representations are likely to lack modality characteristics. Hence, modality-specific multimodal representations are required with the help of various unimodal models to provide unique characteristics of each modality. Modality-specific visual representations $P^v = \{p_1^v, ..., p_{N_s}^v\} \in R^{N_s \times d_p^v}$ is obtained by passing X^v to a pre-trained video model E_V. Meanwhile, modality-specific acoustic representations $P^a =$

$\{p_1^a, ..., p_{N_s}^v\} \in R^{N_s \times d_p^a}$ is encoded from X^a using a pre-trained audio model E_A. A pre-trained language model E_T is employed to extract modality-specific textual representations $P^t = \{p_1^t, ..., p_{N_s}^t\} \in R^{N_s \times d_p^t}$ from X^t.

Text-Guided Nonverbal Representation Selection (TNvRS)

Based on the modality-invariant multimodal representations obtained in the previous step, we aim to select limited shots of critical nonverbal representations with textual guidance, which provides more discriminative information pertinent to VSSR. The core aspect of this module is to effectively model cross-modal interactions, for which we utilize cross-modal attention to represent the relevance of text to nonverbal representations.

Specifically, given textual and nonverbal modality-invariant representations C^t and $C^m, m \in \{v, a\}$, we firstly convert them into the query $Q_t^m \in R^{N_s \times d^c}$ and key $K_m \in R^{N_s \times d^c}$ required in cross-modal attention computation with the help of learnable weights $W_{Q_t^m} \in R^{d^c \times d^c}$, $W_{K_m} \in R^{d^c \times d^c}$ and biases $b_{Q_t^m} \in R^{d^c}$, $b_{K_m} \in R^{d^c}$:

$$Q_t^m = C^t W_{Q_t^m} + b_{Q_t^m}, \tag{1}$$

$$K_m = C^m W_{K_m} + b_{K_m}, \tag{2}$$

where v, a and t represent visual, acoustic and textual modalities, respectively.

Afterward, we take the correlation $A'_{t \to m} \in R^{N_s \times N_s}$ between textual Q and nonverbal K as cross-modal attention. Then we aggregate textual attention of different shots by average to obtain the final attention $A_{t \to m} \in R^{N_s}$ from textual modality to nonverbal modalities:

$$A'_{t \to m} = \text{softmax}\left(\frac{Q_t^m {K_m}^{\mathrm{T}}}{\sqrt{d^c}}\right), \tag{3}$$

$$A_{t \to m} = \frac{1}{N_s} \sum_{i=1}^{N_s} [A'_{t \to m}]_i, \tag{4}$$

where softmax$(\cdot)$ is a softmax normalization, $[]_i$ represents the ith line of matrix.

According to the attention scores, we pick out the first N_m shots of nonverbal representations as critical ones. As a consequent, the remaining nonverbal modality-invariant and -specific representations are $H^m = \{c_{k_1^m}^m, ..., c_{k_{N_m}^m}^m\} \in R^{N_m \times d^c}$ and $Z^m = \{p_{k_1^m}^m, ..., p_{k_{N_m}^m}^m\} \in R^{N_m \times d_p^m}$, where $k_j^m, j \in \{1, ..., N_m\}$ represents the index of the shot whose attention ranking is jth in the original shot sequence S. All modality-invariant and -specific textual representations are remained as $H^t = \{c_1^t, ..., c_{N_t}^t\} \in R^{N_t \times d^c}$ and $Z^t = \{p_1^t, ..., p_{N_t}^t\} \in R^{N_t \times d_p^t}$, where $N_t = N_s$, to preserve abundant available information in text.

Modality-Invariant and -Specific Representation Decoupling (MISD)

Modality-invariant and -specific representations of the same modality may exhibit duplication, as they originate from and describe the same raw data. To avoid the adverse redundancy, it is essential to decouple the modality-invariant and -specific representations to ensure that they contain information of different aspects.

Given modality-specific multimodal representations $Z^m, m \in \{v, a, t\}$, to prepare for the following decoupling, we project them to the same embedding space of modality-invariant representations to obtain $U^m \in R^{N_m \times d^c}$:

$$U^m = E^m\left(Z^m; \theta_e^m\right), \tag{5}$$

where E^m and θ_e^m represent the specific encoder model and parameters, respectively, v, a and t represent visual, acoustic and textual modalities, respectively.

A soft orthogonality metric inspired by (Bousmalis et al. 2016) is calculated to measure the information redundancy between the modality-invariant and -specific representations. The optimization target is the orthogonality average of all three modalities that enforces modality-invariant and -specific representations to achieve non-redundancy:

$$\mathcal{L}_{diff} = \frac{1}{3} \sum_m \left(\|H^m {U^m}^{\mathrm{T}}\|_{1,1}\right), \tag{6}$$

where $\|\cdot\|_{1,1}$ is the "entry-wise" matrix $L_{1,1}$-norm.

To ensure that modality-specific representations retain useful characteristics of each modality without losing important information under the disparity constraint, we apply a soft reconstruction loss for decoding the decoupled modality-specific representations:

$$Z^m\prime = D^m\left(U^m; \theta_d^m\right), \tag{7}$$

$$\mathcal{L}_{recon} = \frac{1}{3} \sum_m \left(\alpha - cos\left(Z^m, Z^m\prime\right)\right), \tag{8}$$

where D^m and θ_d^m represent the specific decoder model and parameters, respectively, α is a distance margin, $cos\left(\cdot\right)$ represents cosine similarity. We apply α to enforce partial, rather than total coincidence between the reconstructed result $Z^m\prime$ and the original input Z^m because of inevitable representation variation during decoupling.

Multimodal Fusion

Both modality-invariant and -specific representations of each modality $H^m \in R^{N_m \times d^c}, m \in \{v, a, t\}$ and $U^m \in R^{N_m \times d^c}$ are aggregated by mean pooling along the temporal dimension to obtain $c^m \in R^{d^c}$ and $u^m \in R^{d^c}$ that represent the whole video. As invariant and specific representations capture different aspects of the same modality, we apply a gate G_m to adaptively merge them, which ensures that both modality-invariant and -specific information are preserved and remain complementary to provide a holistic view for each modality:

$$g^m = \text{sigmoid}(W_{G_m}\left[c^m; u^m\right] + b_{G_m}), \tag{9}$$

$$o^m = (1 - g^m)c^m + g^m u^m, \tag{10}$$

where sigmoid$(\cdot)$ is a sigmoid normalization, $[;]$ represents feature concatenation, W_{G_m} and b_{G_m} are learnable weights and biases of the gate model, respectively, $o^m \in R^{d^c}$ is

the modality-invariant and -specific representation merging result, v, a and t represent visual, acoustic and textual modalities, respectively.

We add all representations to obtain the final multimodal fusion result $o = o^v + o^a + o^t \in R^{d^c}$. The final task prediction $\hat{y}$ is generated by input o to an MLP classifier. We employ the standard cross-entropy loss as task loss $\mathcal{L}_{\text{task}}$:

$$\mathcal{L}_{task} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y_i}), \tag{11}$$

where N is the number of training samples, y_i is the one-hot representation of ground-truth label.

Combing the task loss $\mathcal{L}_{task}$, the decoupling loss $\mathcal{L}_{diff}$ and the reconstruction loss $\mathcal{L}_{recon}$, the final optimization object is computed as:

$$\mathcal{L} = \mathcal{L}_{task} + \lambda_1 \mathcal{L}_{diff} + \lambda_2 \mathcal{L}_{recon}, \tag{12}$$

where λ_1 and λ_2 are weights that determine the contribution of the corresponding module to the overall loss.

Experiments

Datasets and Experimental Settings

Datasets. LVU-VSSR (Wu and Krahenbuhl 2021) is a widely-used VSSR dataset, containing ∼1K conversation videos, each ranging from one to three minutes. There are five categories of speaking styles and each video is assigned with one speaking style label. Consistent with previous work settings, the dataset is divided into train, valid and test sets by the ratio of 5 : 1 : 1. To validate the generalization of TNvE, since the relationship among humans can be revealed by key behaviors, we also evaluate on LVU-VSRR (Wu and Krahenbuhl 2021) dataset, which is a two-person social relationship recognition dataset. There are ∼200 human-centered videos, each of which lasts one to three minutes and relates to one relationship label. The split ratio of train, valid and test sets is 3 : 1 : 1 as prior researches.

Evaluation metrics. Six typical metrics for classification task are used to evaluate model performances: Top-1 Accuracy (Acc), Macro F1-score (F1), Macro Precision (P), Macro Recall (R), Weighted F1-score (WF1) and Weighted Precision (WP). The higher values indicate better performance of all the metrics.

Furthermore, to validate the decoupling ability of MISD module, Earth Movers Distance (EMD) (Rubner, Tomasi, and Guibas 2000) and Proxy A-distance (PAD) (Ben-David et al. 2006) are employed to measure the representation distances. The higher values indicate longer distances.

Implementation details. The shot boundary detector is TransNet V2 (Souček and Lokoč 2020) and the shot number N_s is 15. The automatic speech recognition model is Whisper (Radford et al. 2023). The pre-trained multimodal binding model E_C is ImageBind (Girdhar et al. 2023) and the modality-specific visual, acoustic and textual encoder E_V, E_A and E_T are VideoMAE V2 (Wang et al. 2023c), XLS-R (Babu et al. 2022) and Bert (Jacob Devlin and Toutanova 2019), respectively. Encoder E^m and decoder D^m of each modality in MISD module and the final MLP classifier are all composed of one projection layer. The modality-invariant multimodal representation dimension d^c is 1024 while the modality-specific visual, acoustic and textual representation dimension d_p^v, d_p^a and d_p^t are 768, 1024 and 768, respectively. The distance margin α in reconstruction loss is $\{0.7, 0.2\}$ for LVU-VSSR and LVU-VSRR, respectively. The loss weights λ_1, λ_2 are $\{0.7, 0.8\}$ and $\{0.6, 0.1\}$. The number of selected visual and acoustic shots N_v and N_a are set as $\{1, 1\}$, and $\{5, 6\}$.

In order to preserve the distribution alignment among modality-invariant multimodal representations from being destroyed by the decoupling process, we split the training of TNvE into two stages: 1) we firstly train the TNvRS module via fusing modality-invariant multimodal representations directly for prediction; and 2) we then frozen the parameters of the TNvRS module and train the whole framework including the MISD module. The first stage is trained for $\{100, 50\}$ epoches with batch size 8 using Adam optimizer and a learning rate 4e-6. The second stage is trained for $\{50, 45\}$ epoches with batch size 32 using Adam optimizer and a learning rate 4e-4. All experiments are conducted on a server with 128GB memory, i9-13900K CPU and one RTX4090 GPU of 24GB memory.

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**

Table 1: Comparison results of our TNvE *vs.* different state-of-the-art methods on LVU-VSSR dataset. Here, the best result is in bold.

Comparison with State-of-the-Arts

The comparison results between TNvE and existing VSSR methods are shown in Table 1 and 2, where TNvE consistently outperforms the state-of-the-art methods in all metrics of both datasets, and we can conclude the following observations.

Firstly, TNvE is superior to all unimodal methods, which highlights the importance of multimodal integration and validates the effectiveness of our strategy in multimodal fusion. Secondly, TNvE outperforms all multimodal approaches, which can be contributed to two factors: 1) TNvE effectively captures the critical nonverbal information with

Method	Acc	F1	P	R	WF1	WP
Unimodal						
ObjTrans(Wu and Krahenbuhl 2021)	54.8	28.4	35.0	36.1	41.3	43.1
ViS4mer (Islam and Bertasius 2022)	57.1	40.3	36.4	45.2	51.4	46.7
S5 (Wang et al. 2023b)	67.1	-	-	-	-	-
Multimodal						
MulT (Tsai et al. 2019)	52.4	22.9	17.5	33.3	36.0	28.1
LF-VILA (Sun et al. 2022)	57.1	33.2	51.7	38.9	45.3	57.4
Movie2Scenes (Chen et al. 2023)	71.2	-	-	-	-	-
LMP (Argaw et al. 2023)	69.4	-	-	-	-	-
MA-LLM (He et al. 2024)	57.9	34.8	47.6	40.2	47.9	59.2
LSSD (Singh et al. 2024)	61.0	-	-	-	-	-
TNvE (Ours)	**71.4**	**61.5**	**80.0**	**59.6**	**68.1**	**75.7**

Table 2: Comparison results of our TNvE *vs.* different state-of-the-art methods on LVU-VSRR dataset. Here, the best result is in bold.

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

Table 3: Ablation results of VSSR on LVU-VSSR dataset with different modality combinations w/o *vs.* w/ TNvRS module. Here, v, a and t refers to visual, acoustic and textual modalities, respectively.

textual guide, distinguishing the contribution of different modalities to VSSR, whereas most other methods treat all modalities equally; and 2) even when compared with text-centered methods like Bert-MAG, TNvE achieves superior performance, benefiting from our nonverbal selection strategy and the modality-specific representation incorporation.

Ablation Studies

Text-guided nonverbal representation selection. To demonstrate the effectiveness of text-guided selection, we conduct experiments using selected *vs.* unselected nonverbal representations in TNvRS for VSSR. The comparison results are presented in Table 3, leading to the following insights: 1) text-guided selection can improve unimodal performance, which proves that our cross-modal attention based selection module effectively identifies significant nonverbal representations, reducing noise caused by redundancy in conversation videos; 2) text-guided selection can boost multimodal performances, indicating that multimodal fusion can benefit from enhanced unimodal representation. Correspondingly, unselected nonverbal representations may negatively impact the valid textual representations during fusion, underscoring the importance of nonverbal representation selection; 3) multimodal combinations outperform unimodal variants, especially for that the fusion of all three modalities performs the

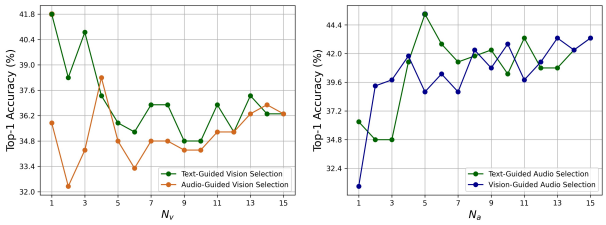

Figure 3: The evaluation curve of unimodal VSSR on LVU-VSSR dataset with different selection number of visual shots N_v and acoustic shots N_a using different modalities as the selection guide.

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

Table 4: Ablation results of VSSR on LVU-VSSR dataset for different variants of MISD module. Here, M-I and M-S refer to modality-invariant and -specific multimodal representations, Diff and Recon refer to the decoupling loss and reconstruction loss, Add, Concat and Gate refer to different modality-invariant and -specific representation fusion strategies.

best, which verifies the necessity of leveraging multimodal information for VSSR. Note here, when only fusing visual and acoustic representations, the performance drops with TNvRS. It is because the text-based cross-modal attention in TNvRS encourages nonverbal representations close to textual contents, leading to a loss of nonverbal information that would otherwise complement each other.

We further conduct experiments using different selection numbers and different modalities as the selection guide to make unimodal predictions. The results of different settings are presented in Figure 3. Compared to nonverbal-guided selection, text-guided selection with selection number $N_v = 1$ and $N_a = 5$ achieves the best performance, demonstrating that text is superior to other modalities to be the selection guide. Note here, curves in Figure 3 exhibit some fluctuations rather than an absolute consistent trend. It is caused by the causal associations among certain shots, for which adding just one shot might introduce noise, while adding two or more shots could be beneficial. As a result, the final selection number is a relative best choice, balancing the trade-off between introducing noise and capturing useful information.

Modality-invariant and -specific representation decoupling. To validate the effectiveness of the MISD module, we conduct ablation studies on the application of modality-specific representations, modality-invariant representations and decoupling losses. In Table 4, the experimental results

Method	v		a		t	
	EMD	PAD	EMD	PAD	EMD	PAD
w/o MISD	1.37	1.94	1.17	1.64	1.41	1.94
w/ MISD	1.39	2.00	1.40	1.72	1.40	1.96

Table 5: Ablation results of the domain distances between invariant and specific representations of each modality on LVU-VSSR dataset w/o *vs.* w/ MISD module. Here, v, a and t represents visual, acoustic and textual representation domains, respectively.

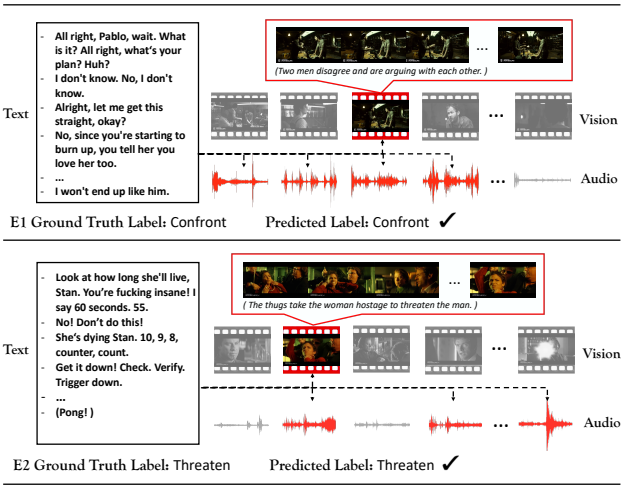

Figure 4: Examples of VSSR results on LVU-VSSR dataset using TNvE. Here, nonverbal shots in red are selected shots guided by text while those in gray are discarded ones.

reveal several key insights: 1) using only modality-specific multimodal representations for TNvRS performs worse than using only modality-invariant ones. This outcome suggests that the distribution heterogeneity among modality-specific representations makes them unsuited for cross-modal attention calculation; 2) simply fusing modality-invariant and -specific multimodal representations without decoupling results in performance drop. This result likely stems from the information duplication between the invariant and specific representations, leading to negative impact on representation fusion; 3) applying the decoupling loss alone makes limited improvement but combining it with the reconstruction loss gains obvious benefit, indicating that the reconstruction process preserves modality-specific representations from losing useful information during decoupling, which demonstrates the effectiveness and validity of our MISD module; 4) fusing modality-invariant and -specific representations via a gate performs better than direct addition or concat, which demonstrates that the designed gate in TNvE facilitates adaptive representation fusion, ensuring that complementary information is retained while minimizing duplication.

In addition, we quantitatively assess the decoupling ability of the MISD module by measuring the domain distances between modality-invariant and -specific representations across different modalities. As shown in Table 5, it is obvious that longer modality-invariant and -specific representation distribution distances are achieved after applying MISD module, which validates the effectiveness of our decoupling strategy.

Visualization

Figure 4 visualizes the text-guided nonverbal selection results of two videos, reflecting the following observations: 1) in the selected visual shot of the top example (E1), two men are depicted arguing, which clearly indicates confrontation and tension in the conversation. In the bottom example (E2), the visual content shows a physical altercation where individuals use violence to coerce a woman, suggesting a threatening situation. TNvE effectively highlights these shots, which clearly embody the corresponding speaking styles, in contrast to other shots that are composed of confusing face sub-window transitions and less relevant visual information; 2) the selected acoustic segments emphasize moments with significant changes in volume and pitch, which reveal the emotional dynamics of the characters. In particular, TNvE captures a gunshot in E2, indicating a violent incident, which strongly suggests that the speaking style is "Threaten" rather than a less intense style like "Explain". This demonstrates that TNvE can prioritize audios that provide crucial context for understanding the target speaking style; and 3) texts in both examples are highly directional to the target speaking styles, which can be estimated with strong confidence. The textual contents are also dense with relevant information, significantly contributing to the identification of the corresponding speaking styles.

Conclusion and Future Work

In this paper, we proposed a text-guided nonverbal enhancement method, TNvE, for VSSR. To deal with the inherent similarity between conversation videos, TNvE selected critical nonverbal information with the guide of text based on modality-invariant multimodal representations. Moreover, to achieve comprehensive multimodal understanding, TNvE incorporated modality-specific multimodal representations and decoupled them from modality-invariant representations. Extensive experiments and ablation studies were conducted to demonstrate the effectiveness of TNvE.

There still exist variety of challenges entailing further consideration. Data limitation restricts the real-world relevance of TNvE, for which VSSR dataset exploration is in great demand. It is difficult to construct a VSSR dataset since it requires not only long-term conversation video collection but also diverse annotations involving exact fine-grained labels. We will devote to contributing a high-quality comprehensive VSSR dataset to advance further research.

Acknowledgments

This work is supported by the National Science Foundation of China (62072232), the Fundamental Research Funds for the Central Universities (021714380026), and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

Aneja, D.; Hoegen, R.; McDuff, D.; and Czerwinski, M. 2021. Understanding conversational and expressive style in a multimodal embodied conversational agent. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 1–10.

Argaw, D. M.; Lee, J.-Y.; Woodson, M.; Kweon, I. S.; and Heilbron, F. C. 2023. Long-range multimodal pretraining for movie understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13392–13403.

Babu, A.; Wang, C.; Tjandra, A.; Lakhotia, K.; Xu, Q.; Goyal, N.; Singh, K.; Von Platen, P.; Saraf, Y.; Pino, J.; et al. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proceedings of the Conference of the International Speech Communication Association*, 2278–2282.

Ben-David, S.; Blitzer, J.; Crammer, K.; and Pereira, F. 2006. Analysis of representations for domain adaptation. In *Proceedings of the International Conference on Neural Information Processing Systems*, 137–144.

Bordwell, D.; Thompson, K.; and Smith, J. 2010. Film art: An introduction. In *McGraw-Hill New York*.

Bousmalis, K.; Trigeorgis, G.; Silberman, N.; Krishnan, D.; and Erhan, D. 2016. Domain separation networks. In *Proceedings of the International Conference on Neural Information Processing Systems*, 343–351.

Chen, S.; Liu, C.-H.; Hao, X.; Nie, X.; Arap, M.; and Hamid, R. 2023. Movies2Scenes: Using movie metadata to learn scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6535–6544.

Fish, E.; Weinbren, J.; and Gilbert, A. 2022. Two-Stream Transformer Architecture for Long Form Video Understanding. In *Proceedings of the British Machine Vision Conference*.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. ImageBind: One Embedding Space To Bind Them All. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15180–15190.

Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the ACM International Conference on Multimedia*, 1122–1131.

He, B.; Li, H.; Jang, Y. K.; Jia, M.; Cao, X.; Shah, A.; Shrivastava, A.; and Lim, S.-N. 2024. MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13504–13514.

Islam, M. M.; and Bertasius, G. 2022. Long movie clip classification with state-space video models. In *Proceedings of the European Conference on Computer Vision*, 87–104.

Jacob Devlin, K. L., Ming-Wei Chang; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.

Li, Y.; Wang, Y.; and Cui, Z. 2023. Decoupled Multimodal Distilling for Emotion Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6631–6640.

Liang, T.; Lin, G.; Feng, L.; Zhang, Y.; and Lv, F. 2021. Attention is not enough: Mitigating the distribution discrepancy in asynchronous multimodal sequence fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8148–8156.

Mai, S.; Hu, H.; and Xing, S. 2020. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 164–172.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the International Conference on Machine Learning*, 28492–28518.

Rahman, W.; Hasan, M. K.; Lee, S.; Bagher Zadeh, A.; Mao, C.; Morency, L.-P.; and Hoque, E. 2020a. Integrating Multimodal Information in Large Pretrained Transformers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2359–2369.

Rahman, W.; Hasan, M. K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.-P.; and Hoque, E. 2020b. Integrating multimodal information in large pretrained transformers. In *Proceedings of the Annual Meeting of Association for Computational Linguistics*, 2359–2369.

Rubner, Y.; Tomasi, C.; and Guibas, L. J. 2000. The earth mover's distance as a metric for image retrieval. In *International Journal of Computer Vision*, volume 40, 99–121.

Singh, N.; Wu, C.-W.; Orife, I.; and Kalayeh, M. 2024. Looking similar sounding different: Leveraging counterfactual cross-modal pairs for audiovisual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26907–26918.

Souček, T.; and Lokoč, J. 2020. Transnet v2: An effective deep network architecture for fast shot transition detection. In *arXiv preprint arXiv:2008.04838*.

Sun, Y.; Xue, H.; Song, R.; Liu, B.; Yang, H.; and Fu, J. 2022. Long-Form Video-Language Pre-Training with Multimodal Temporal Contrastive Learning. In *Proceedings of the International Conference on Neural Information Processing Systems*, 38032–38045.

Sun, Z.; Sarma, P.; Sethares, W.; and Liang, Y. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 8992–8999.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal Transformer

for Unaligned Multimodal Language Sequences. In *Proceedings of the Annual Meeting of Association for Computational Linguistics*, 6558–6569.

Wang, D.; Liu, S.; Wang, Q.; Tian, Y.; He, L.; and Gao, X. 2023a. Cross-modal enhancement network for multimodal sentiment analysis. In *IEEE Transactions on Multimedia*, volume 25, 4909–4921.

Wang, J.; Zhu, W.; Wang, P.; Yu, X.; Liu, L.; Omar, M.; and Hamid, R. 2023b. Selective structured state-spaces for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6387–6397.

Wang, L.; Huang, B.; Zhao, Z.; Tong, Z.; He, Y.; Wang, Y.; Wang, Y.; and Qiao, Y. 2023c. VideoMAE V2: Scaling Video Masked Autoencoders With Dual Masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14549–14560.

Wang, Y.; Shen, Y.; Liu, Z.; Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 7216–7223.

Wu, C.-Y.; and Krahenbuhl, P. 2021. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1884–1894.

Yang, D.; Huang, S.; Kuang, H.; Du, Y.; and Zhang, L. 2022. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the ACM International Conference on Multimedia*, 1642–1651.

Yu, W.; Xu, H.; Yuan, Z.; and Wu, J. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 10790–10797.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1103–1114.

Zhang, B.; Fang, Y.; Yu, F.; Bei, J.; and Ren, T. 2023. MMSF: A Multimodal Sentiment-Fused Method to Recognize Video Speaking Style. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, 289–297.

Zhou, M.; Bai, Y.; Zhang, W.; Yao, T.; Zhao, T.; and Mei, T. 2022. Responsive listening head generation: a benchmark dataset and baseline. In *Proceedings of the European Conference on Computer Vision*, 124–142.