IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 27 April 2023, accepted 22 May 2023, date of publication 26 May 2023, date of current version 1 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3280187

RESEARCH ARTICLE

VideoAdviser: Video Knowledge Distillation for Multimodal Transfer Learning

YANAN WANG[1,2], **(Student Member, IEEE), DONGHUO ZENG**[1], **SHINYA WADA**[1], **AND SATOSHI KURIHARA**[2], **(Member, IEEE)**
[1]Artificial Intelligence Division, KDDI Research Inc., Saitama 356-8502, Japan
[2]School of Science for Open and Environmental Systems, Keio University, Tokyo 223-8522, Japan

Corresponding author: Yanan Wang (wa-yanan@kddi.com)

ABSTRACT Multimodal transfer learning aims to transform pretrained representations of diverse modalities into a common domain space for effective multimodal fusion. However, conventional systems are typically built on the assumption that all modalities exist, and the lack of modalities always leads to poor inference performance. Furthermore, extracting pretrained embeddings for all modalities is computationally inefficient for inference. In this work, to achieve high efficiency-performance multimodal transfer learning, we propose ***VideoAdviser***, a video knowledge distillation method to transfer multimodal knowledge of video-enhanced prompts from a multimodal fundamental model (teacher) to a specific modal fundamental model (student). With an intuition that the best learning performance comes with professional advisers and smart students, we use a CLIP-based teacher model to provide expressive multimodal knowledge supervision signals to a RoBERTa-based student model via optimizing a step-distillation objective loss—first step: the teacher distills multimodal knowledge of video-enhanced prompts from classification logits to a regression logit—second step: the multimodal knowledge is distilled from the regression logit of the teacher to the student. We evaluate our method in two challenging multimodal tasks: video-level sentiment analysis (MOSI and MOSEI datasets) and audio-visual retrieval (VEGAS dataset). The student (requiring only the text modality as input) achieves an MAE score improvement of up to **12.3%** for MOSI and MOSEI. Our method further enhances the state-of-the-art method by **3.4%** mAP score for VEGAS without additional computations for inference. These results suggest the strengths of our method for achieving high efficiency-performance multimodal transfer learning.

INDEX TERMS Multimodal transfer learning, knowledge distillation, fundamental model.

I. INTRODUCTION

Transfer learning is a promising methodology that focuses on transferring pretrained representation domains to nearby target domains [1]. For instance, finetuning a pretrained language model on a small annotated dataset enables high-performance text sentiment analysis [2]. Recent fundamental models on diverse modalities such as language models (*e.g.*, RoBERTa [3], GPT-3 [4]), visual models (*e.g.*, ViT [5]), and multimodal models (*e.g.*, CLIP [6], MEET [7]) have millions of parameters and can provide robust modal representations. With such advancement, multimodal transfer learning aims to transform pretrained representations of diverse modalities into a common domain space for effective multimodal fusion [8], [9]. It has been broadly applied to multimodal tasks such as video-level sentiment analysis [10], [11], [12], and audio/text-video retrieval tasks [13], [14], [15], [16].

Existing works on multimodal transfer learning unify adversarial learning to regularize the embedding distributions between different modalities, leading to effective multimodal fusion [14], [17], [18], [19], [20]. However, conventional systems are typically built on the assumption that all modalities exist, and the lack of modalities always leads to poor inference performance. For instance, vision-language models typically fail to achieve expected performance when given

The associate editor coordinating the review of this manuscript and approving it for publication was Alessandro Floris.

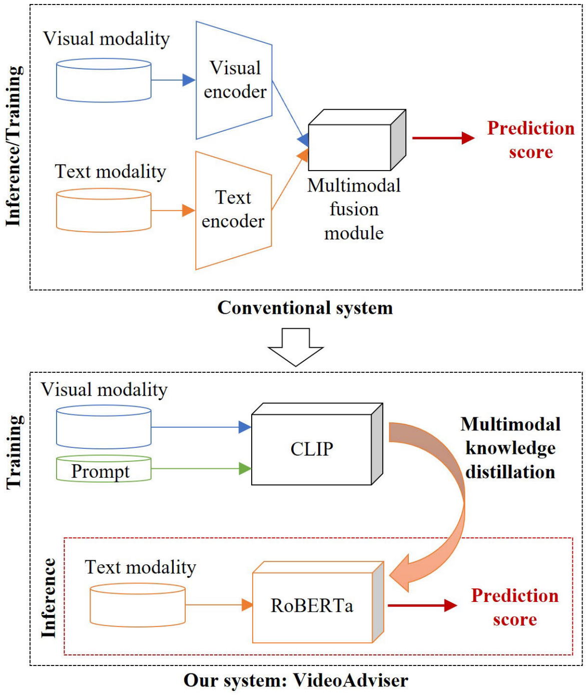


FIGURE 1. A conceptual diagram illustrates the difference between the conventional system and our system: our system focuses on transferring multimodal knowledge from a multimodal fundamental model (*e.g.*, CLIP) to a language fundamental model (*e.g.*, RoBERTa-Large), and requires text only to achieve high efficiency-performance inference. On the other hand, the conventional system focuses on multimodal fusion and requires complex modules (diverse modal encoders and a multimodal fusion module) for inference.

only text data as input. Furthermore, extracting pretrained embeddings for all modalities is computationally inefficient for inference. Therefore, improving robust multimodal transfer learning to achieve high efficiency-performance inference is crucial for practical applications, which motivates this work.

Knowledge distillation (KD) is first proposed for achieving an efficient student model by transforming embedded knowledge in the predicted logits of the teacher model to a smaller student model [21]. Recent works have expanded it to multimodal transfer learning by distilling mutual information from one modality to another [22], [23], [24], [25], [26]. However, these works always need to sacrifice the performance of the teacher model, requiring the teacher model and the student model distributed in neighboring domains (*e.g.*, vision→vison, text→text).

In this paper, with an intuition that the best learning performance comes with professional advisers and smart students, to achieve high efficiency-performance multimodal knowledge distillation, we propose ***VideoAdviser*** shown in Figure 1, a video knowledge distillation method to transfer multimodal knowledge from a strong multimodal fundamental model (teacher) to a powerful specific modal fundamental model (student) via optimizing a step-distillation objective loss. As CLIP is a multimodal fundamental model pretrained with cross-modal contrastive learning on tremendous image-text pairs [6], we employ it as the teacher model to obtain multimodal knowledge of video-enhanced prompts by incorporating the video and text prompt representations. The teacher model utilizes CLIP's visual and text encoders to obtain video and text prompt embeddings without freezing the pretrained weights to preserve multimodal representation space learned by CLIP. By adapting transformer-based modules on these embeddings and extracted frame-level facial expression features, the teacher model acquires expressive multimodal knowledge of video-enhanced prompts by performing video and text prompt representations learning. To sufficiently absorb distilled multimodal knowledge from the teacher model, we employ a large-scale language model RoBERTa [3] as the student model. Since RoBERTa is a transformer-based architecture composed of huge parameters, we finetune its full parameters to leverage RoBERTa's powerful architecture to achieve high-performance student models for inference. In addition, we propose a step-distillation objective loss to distill coarse-fine grained multimodal knowledge to further improve the multimodal knowledge distillation. Motivated by multiscale representation learning enabling the fusion of enriched coarse-fine grained representations [27], [28], we consider that multitask with different target granularities allows the model to acquire representative knowledge at diverse granularities. For instance, classification encourages the model to separate the data point into multiple categorical classes representing an interval of consecutive real values to acquire knowledge at a coarse granularity. In contrast, regression enables the model to distinguish the data point into continuous real values instead of using classes to learn knowledge at a fine granularity. To this end, in the first step, the teacher model distills multimodal knowledge of video-enhanced prompts from classification logits to a regression logit to unify knowledge at both coarse and fine granularity; In the second step, the unified multimodal knowledge is further distilled from the teacher model to the student model.

We evaluate ***VideoAdviser*** in two challenging multimodal tasks: video-level sentiment analysis (MOSI and MOSEI datasets) and audio-visual retrieval (VEGAS dataset). The RoBERTa-based student model requiring only text data as input outperforms the state-of-the-art multimodal model's MAE score by **12.3%** for MOSI and **2.4%** for MOSEI. Our method also enhances the state-of-the-art audio-visual cross-modal model by **3.4%** mAP score for VEGAS without additional computations for inference. Ablation studies further demonstrate that our method is able to improve the state-of-the-art method's MAE score by over **3.0%** with almost half the parameters. These results suggest the strengths of our method for achieving high efficiency-performance multimodal transfer learning.

II. RELATED WORK

1) MULTIMODAL FUNDAMENTAL MODEL

CLIP [6] is a multimodal fundamental model that learns transferable visual models from natural language supervision on a dataset of 400 million (image, text) pairs. It jointly trains an image encoder and a text encoder using contrasting learning objectives to obtain a joint multimodal representation space. Inspired by its remarkable zero-shot generation ability for downstream image tasks, the work [29] proposes XCLIP to expand pretrained CLIP on general video recognition by finetuning it on video data using a video-specific prompting module that enhances the video representation to the text representation. The work [30] utilizes a pretrained CLIP for open-vocabulary object detection by distilling visual knowledge from cropped image regions. In this work, we adapt a pretrained CLIP on distilling multimodal knowledge of video-enhanced prompts from the teacher model to the student model via a step-distillation objective loss.

2) KNOWLEDGE DISTILLATION BASED TRANSFER LEARNING

In addition to achieving a lightweight student model by minimizing the KL divergence between the probabilistic outputs of a teacher and student model [21], recent works on knowledge distillation focus on transferring representational knowledge from a teacher model to a student model [30], [31], [32]. For instance, these works [33], [34] distill linguistic knowledge from a text encoder to a visual encoder by learning the mapping between modal representations. The work [35] utilizes multiple text encoders to perform cross-modal knowledge distillation for stronger text-video retrieval. The work [36] distills expressive text representations from a generation model to the text encoder of CLIP by minimizing text-text feature distance. However, these works mostly focus on knowledge distillation in the common modal domain or show limited performance in the cross-modal domain. In contrast, to achieve expressive knowledge distillation for multimodal transfer learning tasks, we propose a RoBERTa-based student model to improve multimodal knowledge distillation by leveraging its powerful transformer architecture.

3) VIDEO-LEVEL SENTIMENT ANALYSIS TASK

Recent works [2], [10], [11] on video-level sentiment analysis tasks focus on improving modality fusion. The work [18] proposes VAE-Based adversarial learning method to map multimodal representations to a joint domain space for improving the modality fusion process. The work [12] achieves SOTA performance on MOSI [37] and MOSEI [38] dataset by introducing a pretrained modality fusion module that fuses multimodal representation from multi-level textual information by injecting acoustic and visual signals into a text encoder. However, all these works require preprocessed multimodal embeddings as the input which is inefficient for inference. In contrast, we employ a knowledge distillation approach that requires only one specific modality leading to efficient inference.

4) AUDIO-VISUAL RETRIEVAL TASK

Recent works on audio-visual retrieval tasks exploit supervised representation learning methods to generate new features across modalities in a common space [13], [14], [15], [16], [39], [40], [41], [42], such that the audio-visual features can be measured directly. Inspired by the C-CCA [39] that aims at finding linear transformations for each modality, C-DCCA [40] tries to learn non-linear features in the common space by using deep learning methods. Deep learning methods by using rank loss to optimize the predicted distances, such as TNN-C-CCA [13], and CCTL [16] models, which apply triplet losses as the objective functions to achieve better results than other CCA-variant methods. The EICS model [42] learns two different common spaces to capture modality-common and modality-specific features, which achieves the SOTA results so far. In this paper, we enable our method to enhance the extracted audio and visual representations of the SOTA model by distilling multimodal knowledge from a CLIP-based teacher model.

III. PROBLEM SETTING

This work focuses on video-level sentiment analysis and audio-visual retrieval tasks, respectively. For the video-level sentiment analysis task, each data point consists of a video $\boldsymbol{M}$, the cropped sequential face images I, the divided speech text T_{speech}, and the class text T_{class}, our goal is to predict the sentiment intensity $\mathcal{Z}_{pred} \in [-3, 3]$ by giving only speech text T_{speech} for inference. For the audio-visual retrieval task, assume that $\Gamma = \{\gamma_i\}_{i=1}^{N}$ is a video collection, $\gamma_i = \{a_i, v_i\}$, where N indicates the data size, $a_i \in \mathbb{R}^{D1}$ and $v_i \in \mathbb{R}^{D2}$ are audio and visual features from different feature spaces. Our target aims at feeding them into a common space by mapping functions $f(x)$ and $g(x)$ to generate new features $f(a_i)$ and $g(v_i)$. As a result, each query a_i for example will obtain a rank list from another modality based on $query\text{-}v_j (i \neq j)$ similarity.

IV. METHODOLOGY

In this section, we explain our method ***VideoAdviser*** in detail. As shown in Fig. 2, our method consists of a CLIP-based model as the teacher (§ IV-A) and a RoBERTa-based model as the student (§ IV-B). The teacher and student models are jointly trained to achieve knowledge distillation across modalities. The student model enables sentiment intensity prediction by giving only a speech text for inference (§ IV-C). We use $\mathcal{F}(\cdot)$, $\mathcal{V}(\cdot)$, $\mathcal{P}(\cdot)$ and $\mathcal{T}(\cdot)$ to denote the facial expression encoder, visual encoder, prompt encoder, and text encoder.

A. THE CLIP-BASED TEACHER MODEL

1) FACIAL EXPRESSION EMBEDDING

To enhance the visual representations of the teacher model for sentiment intensity prediction, we first use OpenFace [43] to crop face images $\{I_i\}_{i=1}^{T} \in \mathbb{R}^{P^2 \times 3}$ with each of size $P \times P$ pixels from T sampled video frames, then, we extract frame-level facial expression embedding $\boldsymbol{v}^{(f)} \in \mathbb{R}^{T \times D}$ with a facial expression encoder $\mathcal{F}(\cdot)$ [44] that is pretrained on the VGG-Face dataset [45]. Here, $\boldsymbol{v}^{(f)}$ is an 8-dimensional sequential vector of length 64 [$T = 64$, $D = 8$]. More details of the

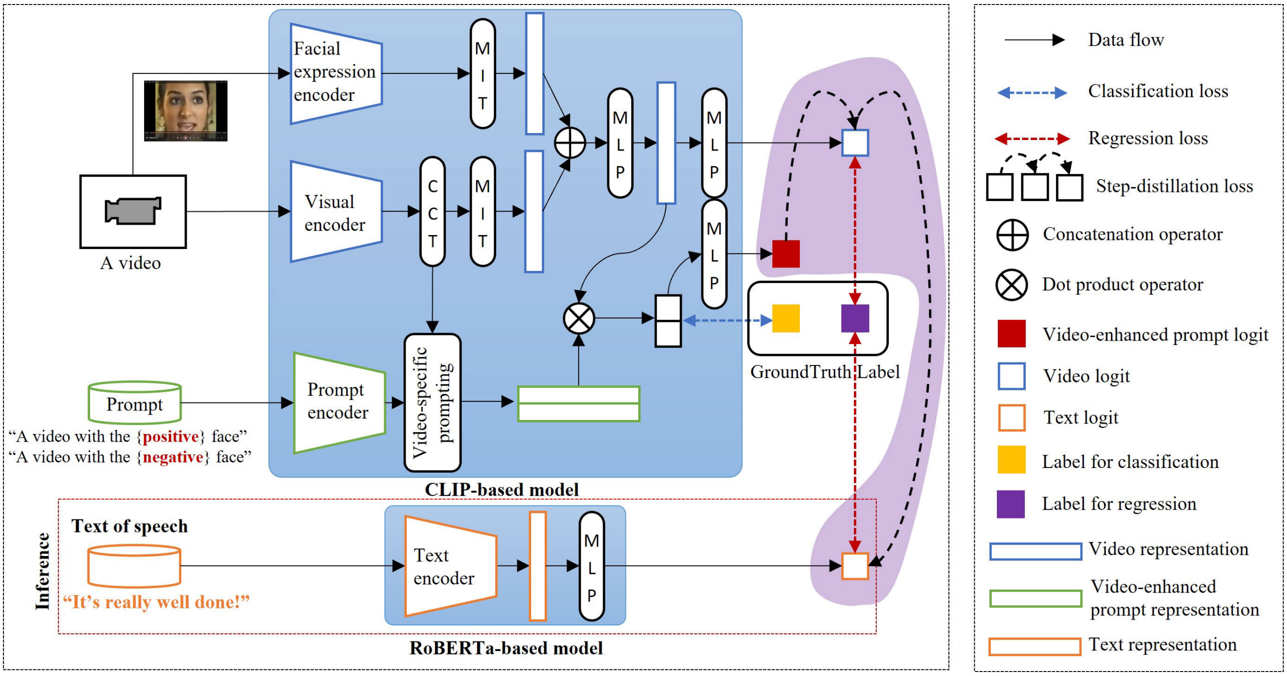


FIGURE 2. Architecture of ***VideoAdviser*** using a CLIP-based model (the teacher) to distill multimodal knowledge of video-enhanced prompts to a RoBERTa-based model (the student): the teacher model utilizes pretrained CLIP's text and visual encoders, and a facial expression encoder to obtain the sentiment class text embedding, the frame-level embedding, and the facial expression embedding. Then, the teacher model employs CCT, MIT, MLP, and a video-specific prompting module, and minimizes a binary sentiment classification loss and a sentiment regression loss. Meanwhile, the student model is finetuned on speech text by minimizing a sentiment regression loss and a step-distillation loss (the region in purple). During inference, the speech text is used to enable sentiment intensity prediction. Here, CCT, MIT, and MLP stand for the cross-frame communication transformer, multi-frame integration transformer, and multi-layer perceptron, respectively.

pretrained model on Albanie's website.[1]

$$\boldsymbol{v}^{(f)} = \mathcal{F}(\{I_i\}_{i=1}^T) \tag{1}$$

2) VISUAL EMBEDDING

To fully transfer the powerful generality of pretrained CLIP [6] from image to video, we freeze the parameters of pretrained CLIP visual encoder $\mathcal{V}(\cdot)$ to obtain frame-level visual embedding $\boldsymbol{v}^{(v)} \in \mathbb{R}^{T\times D}$, where T denotes the number of sampled video frames and D is the dimension of visual embedding. Following [29], given a video clip $M \in \mathbb{R}^{T\times H\times W\times 3}$ of T sampled video frames with $H \times W$ pixels, we use ViT-L/14 [5] to first divide t-th frame into N patches $\{x_{t,i}\}_{i=1}^N \in \mathbb{R}^{P^2\times 3}$, where $t \in T$ and $N = HW/P^2$. Then, the patches $\{x_{t,i}\}_{i=1}^N$ is mapped to $\boldsymbol{v}^{(v)} = \{v_t^{(v)}\}_{t=1}^T$ with a linear transformation $f_m : \mathbb{R}^{P^2\times 3} \to \mathbb{R}^{3P^2\times D}$.

$$\boldsymbol{v}^{(v)} = \mathcal{V}(f_m(\{\boldsymbol{x}_t\}_{t=1}^T)) \tag{2}$$

3) TEXT PROMPT EMBEDDING

We employ the text encoder $\mathcal{P}(\cdot)$ of pretrained CLIP to obtain text prompt embedding $\boldsymbol{v}^{(p)} \in \mathbb{R}^{C\times D}$ of C sentiment classes by giving the sentiment class label $T_{class} \in$ {negative,positive}, where "positive" class includes 0. The text prompt such as "A video with the $\{T_{class}\}$ face" is generated with a text prompt generator f_g and encoded as

$$\boldsymbol{v}^{(p)} = \mathcal{P}(f_g(T_{class})) \tag{3}$$

We employ the cross-frame communication transformer (CCT), multi-frame integration transformer (MIT), and video-specific prompting modules to obtain expressive multimodal sentiment knowledge. The CCT is a multi-layer transformer with cross-frame attention introduced in [29] to enable cross-frame information exchange. It is used to obtain cross-frame visual representations by giving a modified visual embedding $\bar{\boldsymbol{v}}^{(v)} = \{\bar{\boldsymbol{v}}_t^{(v)}\}_{t=1}^T$, where $\bar{\boldsymbol{v}}_t^{(v)} = [x_{class}, v_t^{(v)}] + \boldsymbol{e}_{pos}$. x_{class} is a learnable frame representation and e_{pos} is a position embedding of patches in a frame. The MIT is a normal transformer layer constructed by standard multi-head self-attention and feed-forward networks. Given frame-level embeddings $\boldsymbol{v}^{(f)}$ and $\bar{\boldsymbol{v}}^{(v)}$, we finally obtain the video representation V as follows:

$$V^{(f)} = \text{AvgPool}(\text{MIT}(\boldsymbol{v}^{(f)})) \tag{4}$$

$$V^{(v)} = \text{AvgPool}(\text{MIT}(\text{CCT}(\bar{\boldsymbol{v}}^{(v)}))) \tag{5}$$

$$V = f_v([V^{(f)} || V^{(v)}]) \tag{6}$$

where $f_v : \mathbb{R}^{2\mathcal{D}} \to \mathbb{R}^{\mathcal{D}}$ is a two-layer MLP. AvgPool denotes an average pooling layer. "||" denotes a concatenation operator used to process facial expression-conditioned video representation. We then transform the **video representation** V to the **video logit** (see Fig. 2) with a two-layer MLP.

[1] https://www.robots.ox.ac.uk/ albanie/mcn-models.html

Inspired by [29], the teacher model employs a video-specific prompting module to enhance the prompt embedding with cross-frame visual representations. The video-specific prompting module applies a normal multi-hand attention [46] to obtain the **video-enhanced prompt representation** $\bar{\boldsymbol{v}}^{(p)} \in \mathbb{R}^{C \times D}$ (see Fig. 2) as

$$\bar{\boldsymbol{v}}^{(p)} = \boldsymbol{v}^{(p)} + \text{Multi\_Hand\_Attention}(\text{CCT}(\bar{\boldsymbol{v}}^{(v)})) \tag{7}$$

Then, we compute dot product between video representation V and video-specific prompt representation $\bar{\boldsymbol{v}}^{(p)} = \{\bar{\boldsymbol{v}}_i^{(p)}\}_{i=1}^{C}$ to output the similarity score $\boldsymbol{p} = \{p_i\}_{i=1}^{C}$ with a softmax layer as

$$p_i = \text{softmax}(\bar{\boldsymbol{v}}_i^{(p)} \cdot V) = \frac{\exp(\bar{v}_i^{(p)} \cdot V)}{\sum_{i \in C} \exp(\bar{v}_i^{(p)} \cdot V)} \tag{8}$$

where C indicates the number of sentiment classes. We further transform $\boldsymbol{p}$ into the **video-enhanced prompt logit** (see Fig. 2) with a two-layer MLP.

B. THE RoBERTa-BASED STUDENT MODEL

To leverage the powerful transformer-based architecture of fundamental language models, we structure a RoBERTa-based student model [3] that consists of a text encoder $\mathcal{T}(\cdot)$ and a two-layer MLP. Given the speech text T_{speech}, the student model obtains text representation $V^{(t)}$ with $\mathcal{T}(\cdot)$, and output sentiment intensity $\mathcal{Z}_{pred}$ with the MLP into the **text logit** (see Fig. 2) as

$$\mathcal{Z}_{pred} = \text{logit}(V^{(t)}),\ V^{(t)} = \mathcal{T}(T_{speech}) \tag{9}$$

where $V^{(t)} \in \mathbb{R}^{D}$, and $\text{logit}(\cdot) : \mathbb{R}^{\mathcal{D}} \rightarrow \mathbb{R}^{1}$ indicates the two-layer MLP.

C. TRAINING OBJECTIVES

We simultaneously optimize the teacher and student models by applying mean squared error (MSE) loss to obtain video and text sentiment knowledge. Both teacher and student models minimize the L_2 distance as follows:

$$\mathcal{L}_v^{(r)} = \text{MSE}(\text{logit}(V), l^{(r)}) = \frac{1}{B}\sum_{i=1}^{B} ||\,\text{logit}(V) - l^{(r)}||^2 \tag{10}$$

$$\mathcal{L}_t^{(r)} = \text{MSE}(\mathcal{Z}_{pred}, l^{(r)}) = \frac{1}{B}\sum_{i=1}^{B} ||\mathcal{Z}_{pred} - l^{(r)}||^2 \tag{11}$$

where B indicates batch size, $\mathcal{L}_v^{(r)}$ indicates MSE between the teacher model's video logit and sentiment label $l^{(r)}$, and $\mathcal{L}_t^{(r)}$ indicates MSE between the student model's text logit ($\mathcal{Z}_{pred}$) and $l^{(r)}$. Here, $\text{logit}(V)$ is a two-layer MLP for transforming video representation V into the video logit.

To learn the video-enhanced prompt representation to fuse multimodal knowledge of video and class text, we use the binary sentiment classification label $l^{(c)}$ (see Fig. 3) synthesized from the sentiment label to optimize the teacher model with a cross-entropy loss $\mathcal{L}_v^{(c)}$ as

$$\mathcal{L}_v^{(c)} = -\sum_{i=1}^{C} l_i^{(c)} \log(p_i), \tag{12}$$

We optimize a step-distillation objective loss to achieve multimodal knowledge distillation from the teacher model to the student model. The step-distillation objective loss consists of a **prompt-video distance minimization** $\mathcal{L}_{p \rightarrow v}$ and a **video-text distance minimization** $\mathcal{L}_{v \rightarrow t}$, where $\mathcal{L}_{p \rightarrow v}$ is optimized to align coarse-grained classification knowledge in the video-enhanced prompt logit and fine-grained regression knowledge in the video logit, $\mathcal{L}_{v \rightarrow t}$ is optimized to align knowledge in the video logit of the teacher model and the text logit of the student model. We apply MSE loss to perform the step-distillation as follows:

$$\mathcal{L}_{p \rightarrow v} = \text{MSE}(\text{logit}(\boldsymbol{p}), \text{logit}(V)) \tag{13}$$

$$\mathcal{L}_{v \rightarrow t} = \text{MSE}(\text{logit}(V), \mathcal{Z}_{pred}) \tag{14}$$

where $\text{logit}(\boldsymbol{p})$ indicates the coarse-grained classification knowledge in Eq. 12.

We finally have a joint loss $\mathcal{L}$ for training the teacher and student models end-to-end as

$$\mathcal{L} = \alpha \mathcal{L}_v^{(r)} + \beta \mathcal{L}_t^{(r)} + \gamma \mathcal{L}_v^{(c)} + \delta \mathcal{L}_{p \rightarrow v} + \psi \mathcal{L}_{v \rightarrow t} \tag{15}$$

where α, β, γ, δ, and ψ indicate the importance of each loss value. They are empirically set as $1 : 10 : 1 : 10 : 1$ to keep all loss values on the same scale.

V. EXPERIMENT

In this section, we conducted empirical experiments on video-level sentiment analysis and audio-visual retrieval tasks to demonstrate the high efficiency-performance of our method.

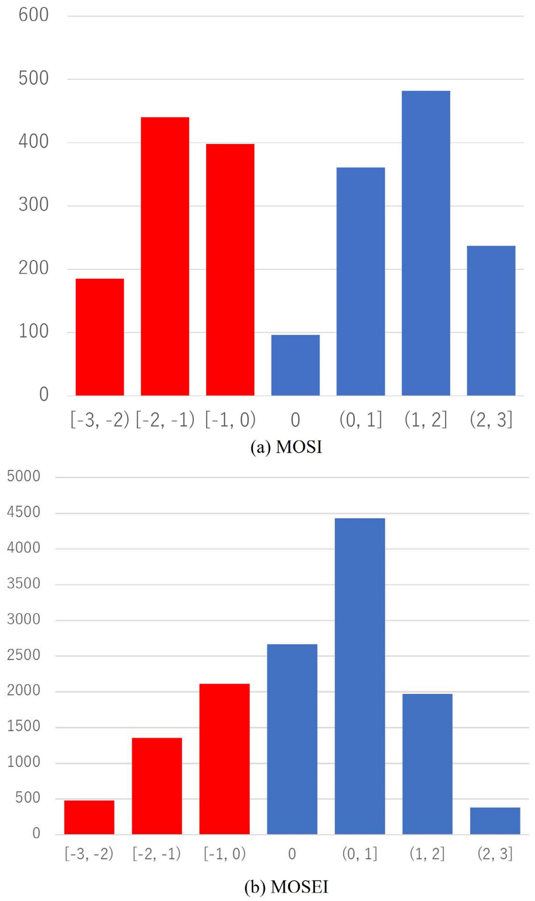
A. DATASET

MOSI [37] and MOSEI [38] are multimodal datasets collected from online video for evaluating video-level sentiment analysis tasks. We show the dataset size in Tab. 1. MOSEI drops the data lacking modalities to fairly evaluate recent modality fusion-based methods [20]. We compared the video segment IDs of each data point for each modality and saved only the data points associated with a common segment ID. The modified MOSEI dataset was found to be more challenging than the original dataset as it lowered the strong baseline MSE score by 4.9% (see Tab. 2). Both datasets are annotated with a Likert scale in the range of $[-3, 3]$, *i.e.*, (-3: highly negative, -2: negative, -1: weakly negative, 0: neutral, +1: weakly positive, +2: positive, +3: highly positive). We further synthesize binary classification label, *i.e.*, ([-3,0): negative, [0,3]: non-negative) used for optimizing the teacher model (§IV-A). The label distribution is illustrated in Fig. 3. MOSEI is imbalanced and over 65% of data is distributed in $[-1, 1]$.

VEGAS dataset [47] is applied for the audio-visual retrieval task, which contains 28,103 videos in total as shown

TABLE 1. Dataset size. MOSEI uses the same dataset as [20].

Dataset	Train	Validation	Test	Total
MOSI [37]	1,284	229	686	2,199
MOSEI [38]	9,473	1,206	2,710	13,389
VEGAS [47]	22,482	-	5,621	28,103



(a) MOSI

(b) MOSEI

FIGURE 3. Label distribution of (a) MOSI and (b) MOSEI. The synthesized binary classification label is illustrated in different colors (the "negative" class in red color and the "non-negative" class in blue color).

in Tab. 1. Each video can be embedded as an audio feature vector and a visual feature vector, and the audio-visual pair shares the same single label. The label represents an audio event (*e.g.*, baby crying) of the human voice or natural sound. The number of label classes is 10, and the length of each audio-visual pair ranges from 2 to 10 seconds.

B. EVALUATION METRIC

We use the mean absolute error (MAE), accuracy (A^7), accuracy (A^2), and weight-$F1$ score for evaluating MOSI and MOSEI. A^7 denotes a 7-class and A^2 denotes a binary accuracy metric. Since MOSI and MOSEI are regression problems, we consider MAE to be the most reasonable metric for fair evaluations. In addition to the binary accuracy reported by most of the previous works, we evaluate the 7-class accuracy as did the SOTA method [12] to eliminate the effect of the data imbalance. For the audio-visual retrieval task, we apply the mean average precision (mAP) as previous works [16], [42] to evaluate our model.

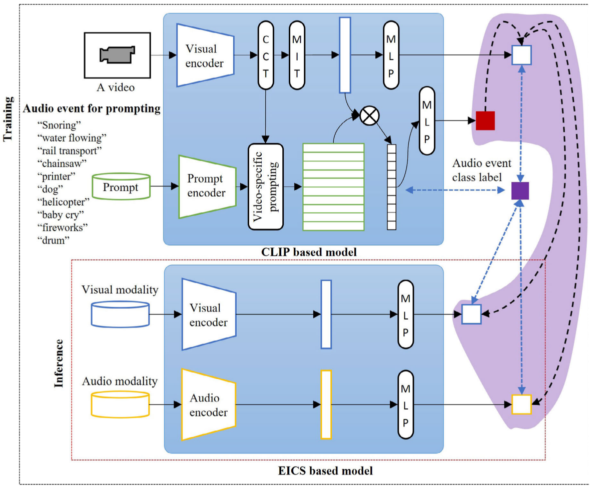


FIGURE 4. Architecture of ***VideoAdviser*** for audio-visual retrieval task using a CLIP-based model (the teacher) to distill multimodal knowledge of video-enhanced prompts to an EICS-based audio-visual model (the student). The teacher model is finetuned for the audio event classification to distill multimodal knowledge to the student model via the step-distillation loss (the region in purple). We adopt 3-layer MLP with 128-dimensional hidden layers.

C. TRAINING SETTING

We train the teacher and the student models simultaneously and use only the student model for inference. The text modality is used for evaluating MOSI and MOSEI. On the other hand, as shown in Fig. 4, we utilize the teacher model to distill multimodal knowledge for both visual and audio encoders of the state-of-the-art model EICS [42] for audio-visual retrieval tasks. Both visual and audio encoders are used as student models to evaluate VEGAS. We show the hyperparameters of ***VideoAdviser*** (§IV) for both tasks in detail in Tab. 3.

D. PERFORMANCE

1) EVALUATION OF VIDEO-LEVEL SENTIMENT ANALYSIS

We compared ***VideoAdviser*** with strong baseline methods on the test set of MOSI and MOSEI in Tab. 2. Compared with the state-of-the-art method UniMSE [12] that utilizes the powerful architecture of a large-scale pretraining model T5 [50] to improve the multimodal fusion by embedding multimodal signals into an auxiliary layer of T5, ***VideoAdviser*** is a multimodal knowledge distillation-based method that distills multimodal knowledge from a multimodal fundamental model CLIP [29] to a language model RoBERTa [3]. UniMSE was trained by integrating the training datasets of MOSI, MOSEI, MELD [51], IEMOCAP [52] and multimodal signals are required for inference. In contrast, our method was trained using the target dataset and requires only text data for inference. ***VideoAdviser*** significantly improves

TABLE 2. Comparison results for MOSI and MOSEI. Our model reduces the state-of-the-art UniMSE's MAE score by 12.3% for MOSI, and VAE-AMDT's MAE by 2.4% for MOSEI. Here, (↓) indicates the lower the MAE, the better the performance, and (↑) indicates the vice-versa. (*) indicates the results produced on the modified MOSEI dataset.

Model	MOSI					MOSEI				
	MAE ↓	A^7 ↑	A^2 ↑	F1 ↑	Corr ↑	MAE ↓	A^7 ↑	A^2 ↑	F1 ↑	Corr ↑
MISA [2]	0.804	-	80.8	80.8	0.764	0.568	-	82.6	82.7	0.717
VAE-AMDT [20]	0.716	-	84.3	84.2	-	0.526*	-	82.8*	**87.5***	-
MAG-BERT [48]	0.712	-	84.2	84.1	0.796	0.539	-	84.7	84.5	-
Self-MM [10]	0.713	-	84.0	84.4	0.798	0.530/0.579*	-	82.8/84.6*	82.5/84.6*	0.765/-
MMM [11]	0.700	46.7	84.1	84.0	0.800	0.526	54.2	82.2	82.7	0.772
UniMSE [12]	0.691	48.7	85.9	85.3	0.809	0.523	54.4	**85.9**	85.8	0.773
VideoAdviser **(ours)**	**0.568**	**51.3**	**87.7**	**87.9**	**0.872**	**0.502***	**54.5***	84.5*	85.0*	**0.810***
Human	0.710	-	85.7	87.5	0.820	-	-	-	-	-

TABLE 3. The hyperparameters for training ***VideoAdviser***. Here, "B" denotes the batch size, "Audio logit" denotes the output of the audio encoder for VEGAS (see Fig. 4).

<table>
<tr><th></th><th>Hyperparameter</th><th>MOSI, MOSEI</th><th>VEGAS</th></tr>
<tr><td rowspan="5">Video</td><td>visual encoder</td><td colspan="2">ViT-L/14</td></tr>
<tr><td>Num. of frames</td><td colspan="2">8</td></tr>
<tr><td>Frame size</td><td colspan="2">224×224</td></tr>
<tr><td>visual embedding size (input)</td><td>(B, 64, 8)</td><td>(B, 1, 10)</td></tr>
<tr><td>Visual hidden layer size</td><td colspan="2">(B, 128)</td></tr>
<tr><td rowspan="3">Prompt</td><td>Prompt encoder</td><td colspan="2">ClipTextModel</td></tr>
<tr><td>Prompt embedding size (input)</td><td colspan="2">(B,77,512)</td></tr>
<tr><td>Prompt hidden layer size</td><td colspan="2">128</td></tr>
<tr><td rowspan="3">Text</td><td>Text encoder</td><td>RoBERTa-large</td><td>-</td></tr>
<tr><td>Text embedding size (input)</td><td>(B,100,1024)</td><td>-</td></tr>
<tr><td>Text hidden layer size</td><td>128</td><td>-</td></tr>
<tr><td rowspan="3">Audio</td><td>Audio encoder</td><td>-</td><td>EICS model</td></tr>
<tr><td>Audio feature size (input)</td><td>-</td><td>10</td></tr>
<tr><td>audio hidden layer size</td><td>-</td><td>128</td></tr>
<tr><td rowspan="4">Output logit</td><td>Video-enhanced prompt logit</td><td>(B, 1)</td><td>(B, 10)</td></tr>
<tr><td>Video logit</td><td>(B, 1)</td><td>(B, 10)</td></tr>
<tr><td>Text logit</td><td>(B, 1)</td><td>-</td></tr>
<tr><td>Audio logit</td><td>-</td><td>(B, 10)</td></tr>
<tr><td rowspan="4">Optimizer</td><td>Method</td><td colspan="2">AdamW [49]</td></tr>
<tr><td>Learning rate</td><td colspan="2">8e-6</td></tr>
<tr><td>Warmup steps</td><td colspan="2">15</td></tr>
<tr><td>Schedular</td><td colspan="2">cosine_schedule_with_warmup</td></tr>
<tr><td rowspan="3">Training</td><td>GPU</td><td colspan="2">GTX 1080 Ti</td></tr>
<tr><td>Batch size</td><td colspan="2">4</td></tr>
<tr><td>Training epochs</td><td colspan="2">100</td></tr>
</table>

TABLE 4. The mAP comparison results with state-of-the-art models for VEGAS. Here, "V" and "A" indicate "Video" and "Audio", respectively.

Model	VEGAS		
	A→V	V→A	Average
Random	0.110	0.109	0.109
BiC-Net [15]	0.680	0.653	0.667
C-CCA [39]	0.711	0.704	0.708
C-DCCA [53]	0.722	0.716	0.719
DCIL [41]	0.726	0.722	0.724
DSCMR [14]	0.732	0.721	0.727
TNN-C-CCA [13]	0.751	0.738	0.745
CCTL [16]	0.766	0.765	0.766
EICS [42]	0.797	0.779	0.788
VideoAdviser **(ours)**	**0.825**	**0.819**	**0.822**

UniMSE's MAE score by **12.3%** for MOSI, and outperforms a strong baseline method VAE-AMDT's MAE score by **2.4%** for MOSEI. As we use the teacher model to offer auxiliary multimodal supervision signals to the student model, by leveraging the strengths of the learned multimodal space of the teacher model and the large-scale parameters of the student model, we think our method is effective for achieving high-performance multimodal knowledge distillation via minimizing the step-distillation objective loss (§IV-C).

2) EVALUATION OF AUDIO-VISUAL RETRIEVAL

We further evaluated our ***VideoAdviser*** on the VEGAS dataset in Tab. 4. Compared to the state-of-the-art method EICS [42] that builds two different common spaces to learn the modality-common and modality-specific features, which achieves an average mAP of 0.788. Our method utilizes the distilled multimodal knowledge to enhance the performance of EICS. As a result, it achieves an average mAP of 0.822 and improves EICS [42] by **3.4%**, suggesting the generality of our method on audio-visual retrieval tasks.

E. EFFICIENCY

By comparing the number of parameters with state-of-the-art models in Fig. 5, our proposed ***VideoAdviser*** requires only a language model as the student that is able to achieve a high efficiency-performance model for inference. The Student (BERT [54]) achieved a compatible MAE score with fewer parameters than previous BERT-based models. Moreover, these models always process visual and audio signals for multimodal fusion, which might require more parameters and increase the computation cost. Compared with the state-of-the-art model UniMSE that uses a pretrained transformer-based language model T5 [50] to perform multimodal fusion, our model, the student (ROBERTa-Base [3]) with nearly half of the parameters reduces MAE score of over **3.0** point, suggesting the high efficiency-performance of our method. ***VideoAdviser*** was further improved over **9.0** point by adopting a RoBERTa-Large model as the student model.

TABLE 5. Efficiency comparison. *VideoAdviser* is able to train a high efficiency-performance student model compared to state-of-the-art methods for inference. The student (RoBERTa-Base) outperforms the SOTA by over 3.0 point with nearly half the parameters.

Model	**Parameters**	**MOSI**
		MAE
BERT-based model		
- MISA [2]	>110M	0.804
- MAG-BERT [48]	>110M	0.712
- Self-MM [10]	>110M	0.713
- MMM [11]	>110M	0.700
T5-based model		
- UniMSE [12]	>231M	0.691
RoBERTa-based model		
- VAE-AMDT [20]	>355M	0.716
***VideoAdviser* (ours)**		
- Student (BERT)	**110M**	0.704
- Student (RoBERTa-Base)	125M	0.660
- Student (RoBERTa-Large)	361M	**0.568**

TABLE 6. Ablation results show the effects of components of the teacher model for multimodal knowledge distillation on MOSI dataset.

Model	**MOSI**			
	MAE	A^7	A^2	F1
VideoAdviser (ours)	**0.568**	**51.3**	87.7	**87.9**
- w/o Facial expression encoder	0.579	50.2	86.8	86.4
- w/o Video-specific prompting	0.570	50.1	**88.1**	87.7

F. ANALYSIS

1) EFFECTIVENESS OF COMPONENTS OF THE TEACHER MODEL

We studied the effects of two core components of the teacher model (Facial expression encoder and video-specific prompting module) in Tab. 6. The results show that these two components help improve the multimodal knowledge distillation and boost the final performance of the student model. We believe that the facial expression encoder provided extra visual knowledge, and the video-specific prompting module further associated visual knowledge with text prompt representations encoded by the prompt encoder.

2) EFFECTIVENESS OF THE STUDENT MODEL

We studied the effects of ***VideoAdviser*** on different student models in Tab. 7. We select two language models (BERT and RoBERTa) that have frequently been used in recent works [2], [10], [11], [20], [48]. By comparing the performance of language models with and without adopting a teacher model, the results demonstrate that our method improves a general language model's MAE score by over **6.0** point on average, suggesting the efficacy and generality of our method with different student models. We consider that the teacher model offers auxiliary multimodal supervision to the student model during training, the language model-based students are able to learn multimodal knowledge from the teacher with their large-scale parameters.

TABLE 7. Effects in different student models. Our method improves the MAE score of pretrained language models by over 6.0 point on average.

Model	**MOSI**		
	MAE	A^2	F1
Teacher (CLIP-based model)	-	57.3	-
BERT w/o teacher	0.753	84.1	83.6
Student (BERT)	0.704	84.7	83.8
RoBERTa-Base w/o teacher	0.719	84.6	84.3
Student (RoBERTa-Base)	0.660	85.4	84.6
RoBERTa-Large w/o teacher	0.660	87.3	87.3
Student (RoBERTa-Large)	**0.568**	**87.7**	**87.9**

TABLE 8. Ablation results show the effects of step-distillation on audio and video modalities for VEGAS. Here, "w/ video distillation" indicates that the step-distillation is only adopted for the visual modality of the student model, "w/ audio distillation" indicates the other side, and "w/ audio and video distillation" indicates both sides (see Fig. 4).

Model	**VEGAS**		
	A→V	V→A	Average
baseline (EICS [42])	0.797	0.779	0.788
***VideoAdviser* (ours)**			
-w/ video distillation	0.794	0.810	0.802
-w/ audio distillation	0.791	0.815	0.803
-w/ (audio and video) distillation	**0.825**	**0.819**	**0.822**

We further trained a student model by freezing pretrained parameters, which dramatically dropped the MAE score from 0.568 to 1.478. This result makes us believe that in order to achieve expressive multimodal knowledge distillation across modalities, it is essential to finetune full parameters to leverage the strengths of large-scale pretrained models with powerful representational learning capabilities.

3) MODALITY EFFECTIVENESS

To confirm the robustness of ***VideoAdviser*** in multimodal knowledge distillation not only for text modality but also for diverse modalities such as visual and audio modalities, we respectively studied the effects on visual and audio modalities for audio-visual retrieval tasks. As the results indicated in Tab. 8, the proposed step-distillation works for both modalities by boosting the baseline EICS model by over 1% mAP score. By associating both sides, we finally improved the baseline by 3.4%.

4) EFFECTIVENESS OF DATASET SIZE

In general, the larger the dataset, the better the performance. We trained ***VideoAdviser*** with a combination of the MOSI and MOSEI datasets to see if we can further improve the performance. As the results indicated in Tab. 9, The model performs much better than those trained on individual datasets and suggests the efficacy of our approach for different dataset sizes.

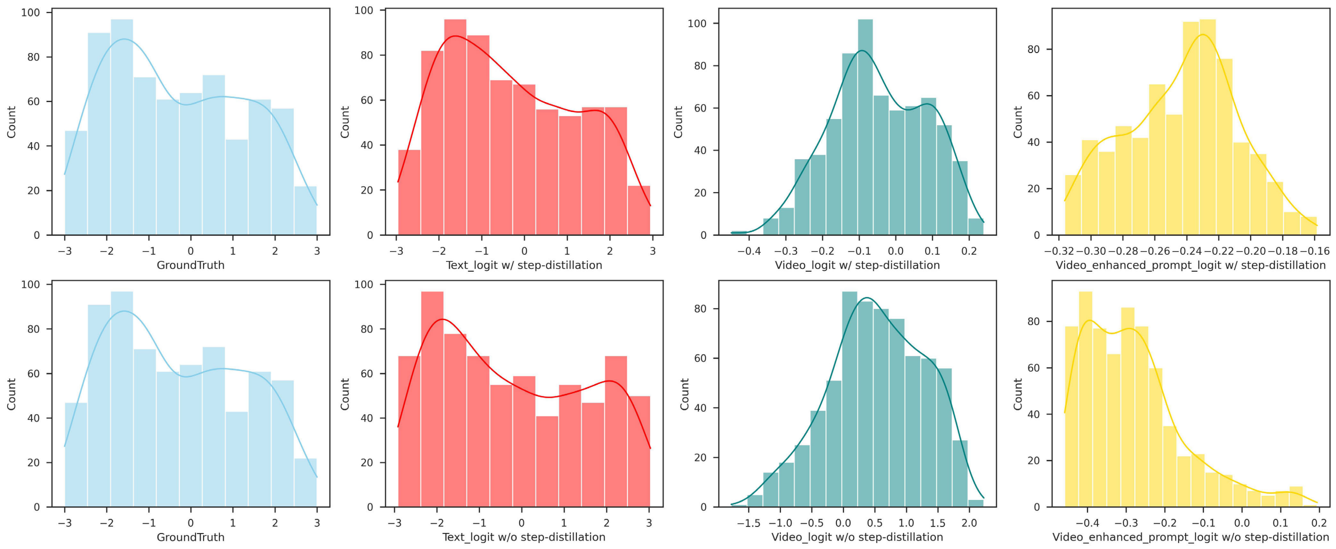


FIGURE 5. Visualization of logistic knowledge distribution with and without the step-distillation objective loss. The top row plots the histograms of logit by applying the step-distillation, and the bottom row indicates the vice-versa. The groudTruth indicates the label distribution, and text_logit indicates the predicted regression score of the student model. Our method using the step-distillation (the top) demonstrates a distribution of regression scores close to the groundTruth, affected by the knowledge distribution of the "video_logit" and "video_enhanced_prompt_logit".

TABLE 9. Results of ***VideoAdviser*** trained with a combination of MOSI and MOSEI datasets. The model performs much better for both the MOSI and MOSEI test sets. Here, (*) denotes the result of the model trained on the individual dataset.

Test dataset	MAE	A^7	A^2
MOSI	**0.546** (0.568)	**51.3** (51.3)	**88.5** (87.7)
MOSEI	**0.491** (0.502)	**55.6** (54.5)	84.2 (**84.5**)
MOSI+MOSEI	0.502	54.79	85.05

TABLE 10. Ablation results show the effects of the proposed step-distillation loss for MOSI.

Model	**MOSI**			
	MAE	A^7	A^2	F1
CRD [31]	0.617	48.8	86.3	85.9
***VideoAdviser* (ours)**				
- w/o step-distillation	0.660	45.5	87.3	87.3
- w/o step-distillation:step1	0.618	49.0	86.5	86.3
- w/ step-distillation	**0.568**	**51.3**	87.7	**87.9**

5) EFFECTIVENESS OF THE STEP-DISTILLATION LOSS

We ablatively studied the effects of our proposed step-distillation loss for multimodal knowledge distillation in Tab. 10. Without the first step—distilling multimodal knowledge from a video-enhanced prompt logit to a video logit (see Fig. 2), the learned multimodal space of CLIP cannot be passed to the student model via the video logit, resulting poor student model performance. On the other hand, it improves the regular language model (w/o step-distillation) **4.2%** MAE score and suggests the effectiveness of the second step—distilling the knowledge of the video logit from the teacher model to the student model. Moreover, by optimizing the first and second steps, our proposed method outperforms a cutting-edge contrastive representation distillation method (CRD) [31] that proposed a contrastive-based objective for transferring knowledge between deep networks. Compared to the CRD which is designed to model mutual information across dimensions of the knowledge representations, Our proposed step-distillation applies MSE to mapping mutual information across modalities via one-dimensional logits (*i.e.*, video-enhanced prompt logit, video logit, and text logit). Our method performs better than CRD in transferring regression information for multimodal knowledge distillation.

In addition, we show comparison results of the proposed step-distillation loss with three widely-known distillation function KD [21], FitNet [55] and PKT [56] in Tab. 11. KD and PKT are proposed to minimize the KL divergence between the probabilistic outputs of a teacher and student model. On the other hand, FitNet and our step-distillation aim at minimizing the L_2 distance for knowledge distillation. Compared to KD, FitNet and PKT are one-step distillation loss functions, whereas our step-distillation performs two-step distillation, with the aim of transferring multimodal knowledge across multiple scales. To achieve a fair comparison, we adapted these three approaches to our problem setting of two-step distillation. As the results indicated in Tab. 11, the step-distillation outperforms other approaches and suggests its efficacy on multimodal knowledge distillation. We noted that the PKT-based two-step distillation achieves a compatible score with ours. We consider that audio-visual tasks focus on distilling multimodal knowledge of categorical audio events rather than fine-grained regressional knowledge so that transferring probabilistic knowledge of each category can also work well. Compared to KD which utilized the softmax function to obtain probabilistic

TABLE 11. Comparison results between widely-known knowledge distillation loss and the proposed step-distillation loss for VEGAS.

Model	VEGAS		
	A→V	V→A	Average
KD [21]	0.783	0.612	0.701
FitNet [55]	0.803	0.781	0.792
PKT [56]	0.824	0.807	0.816
step-distillation (ours)	**0.825**	**0.819**	**0.822**

TABLE 12. ASO scores of models with different distillation objectives studied in Sec. V-F. For "***VideoAdviser*** (ours) → baseline", $\epsilon_{\min} = 0$ indicates that ***VideoAdviser*** (ours) consistently outperform baseline. Here, the baseline denotes ***VideoAdviser*** (ours) w/o step-distillation.

Model	ASO score ($\epsilon_{\min}$)
VideoAdviser (ours) → baseline	0
VideoAdviser (ours) → CRD	0
CRD → baseline	0.02

knowledge, PKT adopted the cos-similarity function to better obtain dimension-level correlation with the probabilistic knowledge.

We further illustrate the logistic knowledge distribution with and without the step-distillation loss in Fig. 5. Compared to the "Text_logit w/o step-distillation" that plots the histogram of regression scores without performing the step-distillation, "Text_logit w/ step-distillation" is close to the groundTruth label distribution. Especially the distribution in the range of $[-1, 1]$ is strongly affected by the teacher model. Because the "Video_logit w/o step-distillation" distributes in the range of $[-1.5, 2]$ and the "Video_enhanced_prompt_logit w/o step-distillation" distributes in the range of $[-0.4, 0.2]$, by performing the step-distillation, the predicted regression score produced by the student model can be affected by the gap of these different distributions, and demonstrate that our proposed step-distillation is effective for multimodal knowledge distillation.

G. SIGNIFICANCE TESTING

We tested the stability of the performance improvement by ***VideoAdviser*** using the Almost Stochastic Order test (ASO) [57], [58] as implemented by [59]. We compared three models, ***VideoAdviser*** (ours), ***VideoAdviser*** w/o step-distillation (baseline), and CRD based on five random seeds each using ASO with a confidence level of $\alpha = 0.05$. ASO computes a score ($\epsilon_{\min}$) indicated in Tab. 12 to represent how far the first model is from being significantly better with respect to the second. $\epsilon_{\min} = 0$ represents truly stochastic dominance and $\epsilon_{\min} < 0.5$ represents almost stochastic dominance.

VI. CONCLUSION

We proposed a novel multimodal knowledge distillation method, ***VideoAdviser***, which leverages the strengths of learned multimodal space of the CLIP-based teacher model and large-scale parameters of the RoBERTa-based student model to perform multimodal knowledge transfer by optimizing a step-distillation objective loss. In the evaluation of two multimodal tasks, our method significantly outperforms SoTA methods up to **12.3%** MAE score with a single modal encoder used in inference for video-level sentiment analysis, and **3.4%** mAP for audio-visual retrieval tasks, suggesting its strengths in high efficiency-performance. Ablation studies further demonstrate the efficacy of our proposed step-distillation objective loss in improving multimodal knowledge distillation. In the next step, we will adapt meta-learning to further explore the capability of multimodal transfer learning in a few-shot setting.

REFERENCES

[1] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.

[2] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and -specific representations for multimodal sentiment analysis," 2020, *arXiv:2005.03545*.

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[4] T. B. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1877–1901.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. ICML*, 2021.

[7] L. Nie, L. Qu, D. Meng, M. Zhang, Q. Tian, and A. D. Bimbo, "Search-oriented micro-video captioning," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 3234–3243.

[8] L. Zhen, P. Hu, Peng, R. S. M. Goh, and J. T. Zhou, "Deep multimodal transfer learning for cross-modal retrieval," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 798–810, Feb. 2022.

[9] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 292–301.

[10] W. Yu, H. Xu, Y. Ziqi, and W. Jiele, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc. AAAI*, 2021.

[11] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 9180–9192.

[12] G. Hu, T.-E. Lin, Y. Zhao, G. Lu, Y. Wu, and Y. Li, "UniMSE: Towards unified multimodal sentiment analysis and emotion recognition," in *Proc. EMNLP*, 2022, pp. 7837–7851.

[13] D. Zeng, Y. Yu, and K. Oyama, "Deep triplet neural networks with cluster-CCA for audio-visual cross-modal retrieval," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 3, pp. 1–23, Aug. 2020.

[14] L. Zhen, P. Hu, X. Wang, and D. Peng, "Deep supervised cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10386–10395.

[15] N. Han, J. Chen, C. Shi, Y. Zeng, G. Xiao, and H. Chen, "BiC-Net: Learning efficient spatio-temporal relation for text-video retrieval," 2021, *arXiv:2110.15609*.

[16] D. Zeng, Y. Wang, J. Wu, and K. Ikeda, "Complete cross-triplet loss in label space for audio-visual cross-modal retrieval," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2022, pp. 1–9.

[17] L. He, X. Xu, H. Lu, Y. Yang, F. Shen, and H. T. Shen, "Unsupervised cross-modal retrieval through adversarial learning," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 1153–1158.

[18] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. MM*, 2017, pp. 154–162.

[19] J. Zhang, Y. Peng, and M. Yuan, ''Unsupervised generative adversarial cross-modal hashing,'' in *Proc. AAAI*, 2017.

[20] Y. Wang, J. Wu, K. Furumai, S. Wada, and S. Kurihara, ''VAE-based adversarial multimodal domain transfer for video-level sentiment analysis,'' *IEEE Access*, vol. 10, pp. 51315–51324, 2022.

[21] G. Hinton, O. Vinyals, and J. Dean, ''Distilling the knowledge in a neural network,'' 2015, *arXiv:1503.02531*.

[22] S. Gupta, J. Hoffman, and J. Malik, ''Cross modal distillation for supervision transfer,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2827–2836.

[23] F. M. Thoker and J. Gall, ''Cross-modal knowledge distillation for action recognition,'' in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 6–10.

[24] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, ''TinyBERT: Distilling BERT for natural language understanding,'' in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*, 2020, pp. 4163–4174.

[25] L. Hou, Z. Huang, L. Shang, X. Jiang, X. Chen, and Q. Liu, ''DynaBERT: Dynamic BERT with adaptive width and depth,'' in *Proc. NeurIPS*, 2020, pp. 9782–9793.

[26] B. Pan, H. Cai, D. Huang, K. Lee, A. Gaidon, E. Adeli, and J. C. Niebles, ''Spatio-temporal graph for video captioning with knowledge distillation,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10867–10876.

[27] Z. Li, Y. Nie, K. Han, J. Guo, L. Xie, and Y. Wang, ''A transformer-based object detector with coarse-fine crossing representations,'' in *Proc. NeurIPS*, 2022.

[28] L. Jiao, J. Gao, X. Liu, F. Liu, S. Yang, and B. Hou, ''Multiscale representation learning for image classification: A survey,'' *IEEE Trans. Artif. Intell.*, vol. 4, no. 1, pp. 23–43, Feb. 2023.

[29] B. Ni, H. Peng, M. Chen, S. Zhang, G. Meng, J. Fu, S. Xiang, and H. Ling, ''Expanding language-image pretrained models for general video recognition,'' in *Proc. ECCV* 2022.

[30] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, ''Open-vocabulary object detection via vision and language knowledge distillation,'' 2021, *arXiv:2104.13921*.

[31] L. Chen, D. Wang, Z. Gan, J. Liu, R. Henao, and L. Carin, ''Wasserstein contrastive representation distillation,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1–6.

[32] L. Wang and K. Yoon, ''Knowledge distillation and student–teacher learning for visual intelligence: A review and new outlooks,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3048–3068, Jun. 2022.

[33] Y. Wang, J. Wu, P. Heracleous, S. Wada, R. Kimura, and S. Kurihara, ''Implicit knowledge injectable cross attention audiovisual model for group emotion recognition,'' in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 827–834.

[34] Y. Wang, J. Wu, J. Huang, G. Hattori, Y. Takishima, S. Wada, R. Kimura, J. Chen, and S. Kurihara, ''LDNN: Linguistic knowledge injectable deep neural network for group cohesiveness understanding,'' in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2020, pp. 343–350.

[35] I. Croitoru, S. Bogolin, M. Leordeanu, H. Jin, A. Zisserman, S. Albanie, and Y. Liu, ''TeachText: CrossModal generalized distillation for text-video retrieval,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11563–11573.

[36] W. Dai, L. Hou, L. Shang, X. Jiang, Q. Liu, and P. Fung, ''Enabling multimodal generation on CLIP via vision-language knowledge distillation,'' in *Proc. Findings Assoc. Comput. Linguistics, ACL*, 2022, pp. 2383–2395.

[37] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, ''MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos,'' 2016, *arXiv:1606.06259*.

[38] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, ''Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph,'' in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.

[39] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal, ''Cluster canonical correlation analysis,'' in *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS)*, vol. 33, 2014, pp. 823–831.

[40] D. Zeng, Y. Yu, and K. Oyama, ''Audio-visual embedding for cross-modal music video retrieval through supervised deep CCA,'' in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2018, pp. 143–150.

[41] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen, ''Dual-path convolutional image-text embeddings with instance loss,'' *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 2, pp. 1–23, May 2020.

[42] D. Zeng, J. Wu, G. Hattori, R. Xu, and Y. Yu, ''Learning explicit and implicit dual common subspaces for audio-visual cross-modal retrieval,'' *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 2s, pp. 1–23, Jun. 2023.

[43] T. Baltrušaitis, P. Robinson, and L. Morency, ''OpenFace: An open source facial behavior analysis toolkit,'' in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–10.

[44] S. Albanie and A. Vedaldi, ''Learning grimaces by watching TV,'' in *Proc. BMVC*, 2016.

[45] O. M. Parkhi, A. Vedaldi, and A. Zisserman, ''Deep face recognition,'' in *Proc. BMVC*, 2015.

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, ''Attention is all you need,'' in *Proc. NeurIPS*, 2017.

[47] Y. Zhou, Z. Wang, C. Fang, T. Bui, and T. L. Berg, ''Visual to sound: Generating natural sound for videos in the wild,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3550–3558.

[48] W. Rahman, M. K. Hasan, S. Lee, A. B. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, ''Integrating multimodal information in large pretrained transformers,'' in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2359–2369.

[49] D. P. Kingma and J. Ba, ''Adam: A method for stochastic optimization,'' 2015, *arXiv:1412.6980*.

[50] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, ''Exploring the limits of transfer learning with a unified text-to-text transformer,'' *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 1–15, 2020.

[51] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, ''MELD: A multimodal multi-party dataset for emotion recognition in conversations,'' in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 527–536.

[52] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, ''IEMOCAP: Interactive emotional dyadic motion capture database,'' *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[53] Y. Yu, S. Tang, K. Aizawa, and A. Aizawa, ''Category-based deep CCA for fine-grained venue discovery from multimodal data,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 4, pp. 1250–1258, Apr. 2019.

[54] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, ''BERT: Pre-training of deep bidirectional transformers for language understanding,'' in *Proc. NAACL*, 2019, pp. 4171–4186.

[55] A. Romero, N. Ballas, S. Ebrahimi Kahou, A. Chassang, C. Gatta, and Y. Bengio, ''FitNets: Hints for thin deep nets,'' 2014, *arXiv:1412.6550*.

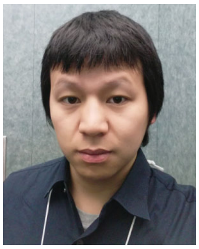
[56] N. Passalis and A. Tefas, ''Learning deep representations with probabilistic knowledge transfer,'' in *Proc. ECCV*, 2018.

[57] E. Del Barrio, J. A. Cuesta-Albertos, and C. Matrán, ''An optimal transportation approach for assessing almost stochastic order,'' in *The Mathematics of the Uncertain*. Cham, Switzerland: Springer, 2018, pp. 33–44.

[58] R. Dror, S. Shlomov, and R. Reichart, ''Deep dominance—How to properly compare deep neural models,'' in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 2773–2785.

[59] D. Ulmer, C. Hardmeier, and J. Frellsen, ''Deep-significance–easy and meaningful statistical significance testing in the age of neural networks,'' 2022, *arXiv:2204.06815*.

YANAN WANG (Student Member, IEEE) received the B.S. degree in engineering from Aoyama Gakuin University, in 2015, and the M.S. degree in engineering from The University of Electro-Communications, Japan, in 2017. He is currently pursuing the Ph.D. degree in engineering with Keio University, Japan. He is an Associate Research Engineer in multimodal modeling topics with KDDI Research Inc. His research interests include multimodal representation learning, emotion recognition, knowledge graph, and graph representation learning. He is a Student Member of JSAI. He is a regular member of IEICE and the Editorial Committee of IEICE Human Communication Group.



DONGHUO ZENG received the M.Sc. degree from the School of Computer Science and Technology, HIT, China, in 2017, and the Ph.D. degree from the National Institute of Informatics, SOKENDAI, Tokyo, Japan, in 2020. He is an AI Researcher of KDDI Research Inc., Tokyo/Saitama, Japan. His research interests include audio-visual learning and video captioning.

SHINYA WADA received the B.E. and M.E. degrees from Kyushu University, in 2005 and 2007, respectively. He is currently a Senior Manager of the Multimodal Modeling Laboratory, KDDI Research Inc. His research interests include multimodal representation learning, human activity recognition, and time-series analysis. He is a member of IEICE.

SATOSHI KURIHARA (Member, IEEE) received the B.E., M.E., and Ph.D. degrees in computer science from Keio University, Tokyo, Japan, in 1990, 1992, and 2000, respectively. In 1992, he joined the Basic Research Division, Nippon Telegraph and Telephone Corporation (NTT). In 2004, he joined the Graduate School of Information Science and Technology/the Institute of Scientific and Industrial Research, Osaka University, Osaka, Japan. In 2013, he joined the Graduate School of Information Systems, The University of Electro-Communications. In 2018, he joined the Faculty of Science and Technology, Keio University, as a Professor. Since April 2021, he has been the Director of the Center of Advanced Research for Human-AI Symbiosis Society. His current research interests include multi-agent systems, ubiquitous computing, and complex network research. He is a member of ACM, AAAI, the Information Processing Society of Japan (IPSJ), the Japan Society of Artificial Intelligence (JSAI), the Institute of Electronics, Information, and Communication Engineers (IEICE), the Society for Economic Science with Heterogeneous Interacting Agents (ESHIA), and the Japan Society of Software Science and Technology (JSSST).