Multimodal Cross-Lingual Summarization for Videos: A Revisit in Knowledge Distillation Induced Triple-Stage Training Method

Nayu Liu, Kaiwen Wei, *Member, IEEE*, Yong Yang, *Senior Member, IEEE*, Jianhua Tao, *Senior Member, IEEE*, Xian Sun, *Senior Member, IEEE*, Fanglong Yao, *Member, IEEE*, Hongfeng Yu, Li Jin, Zhao Lv, *Member, IEEE*, and Cunhang Fan, *Member, IEEE*

***Abstract*—Multimodal summarization (MS) for videos aims to generate summaries from multi-source information (e.g., video and text transcript), showing promising progress recently. However, existing works are limited to monolingual scenarios, neglecting non-native viewers' needs to understand videos in other languages. It stimulates us to introduce multimodal cross-lingual summarization for videos (MCLS), which aims to generate cross-lingual summaries from multimodal input of videos. Considering the challenge of high annotation cost and resource constraints in MCLS, we propose a knowledge distillation (KD) induced triple-stage training method to assist MCLS by transferring knowledge from abundant monolingual MS data to those data with insufficient volumes. In the triple-stage training method, a video-guided dual fusion network (VDF) is designed as the backbone network to integrate multimodal and cross-lingual information through diverse fusion strategies in the encoder and decoder; What's more, we propose two cross-lingual knowledge distillation strategies: adaptive pooling distillation and language-adaptive warping distillation (LAWD), designed for encoder-level and vocab-level distillation objects to facilitate effective knowledge transfer across cross-lingual sequences of varying lengths between MS and MCLS models. Specifically, to tackle lingual sequences of varying lengths between MS and MCLS models. Specifically, to tackle the challenge of unequal length of parallel cross-language sequences in KD, LAWD can directly conduct cross-language distillation while keeping the language feature shape unchanged to reduce potential information loss. We meticulously annotated the How2-MCLS dataset based on the How2 dataset to simulate MCLS scenarios. Experimental results show that the proposed method achieves competitive performance compared to strong baselines, and can bring substantial performance improvements to MCLS models by transferring knowledge from the MS model.**

***Index Terms*—Natural language Processing, multimodal summarization (MS), cross-lingual summarization, abstractive summarization, knowledge distillation (KD), cross-lingual alignment.**

I. Introduction

MULTIMODAL summarization (MS) for videos aims at integrating multimodal information such as videos and text transcripts to generate text summaries. With the rapid growth of videos on the Internet, this task has attracted much interest from the communities and has shown its potential in recent years. It benefits users from better understanding and accessing those lengthy and intricate videos.

Since Libovicky et al. and Palaskar et et al. [1], [2] introduced the multimodal summarization for open-domain videos, former state-of-the-art multimodal summarization methods [3], [4], [5], [6], [7] have achieved great outcomes following the predefined task. While, existing methods are all conducted in monolingual scenarios. In practical applications, for non-native video viewers, they desire some native language summaries to better understand the contents of the videos in other languages. To the best of our knowledge, no research has addressed the problem of generating native language summaries of cross-lingual videos for non-native viewers.

To assist non-native viewers, we introduce the study of multimodal cross-lingual summarization for videos (MCLS). As illustrated in the example from Fig. 1, MCLS seeks to generate summaries in the target language to reflect the salient video contents based on the videos and their transcription text in the source language.

In MCLS, it is not merely necessary to integrate lengthy multi-modal information and different languages to generate summaries compared to MS and cross-lingual summarization

Received 16 November 2023; revised 19 June 2024; accepted 10 August 2024. Date of publication 22 August 2024; date of current version 5 November 2024. This work is supported in part by the STI2030—Major Projects under Grant 2021ZD0201500, in part by the National Natural Science Foundation of China (NSFC) under Grant 62201002 , Grant 62406223 , and Grant 62176029, in part by the Distinguished Youth Foundation of Anhui Scientific Committee under Grant 2208085J05, in part by the Special Fund for Key Program of Science and Technology of Anhui Province under Grant 202203a07020008, and in part by the Open Fund of Key Labora tory of Flight Techniques and Flight Safety, CACC under Grant FZ2022KF15. Recommended for acceptance by L. Sigal. *(Nayu Liu and Kaiwen Wei contributed equally to this work.) (Corresponding author: Cunhang Fan.)*

Nayu Liu and Yong Yang are with the School of Computer Science and Technology, Tiangong University, Tianjin 300387, China (e-mail: nayuliu@tiangong.edu.cn; greatyangy@126.com).

Kaiwen Wei is with the College of Computer Science, Chongqing University, Chongqing 400044, China (e-mail: weikaiwen@cqu.edu.cn).

Xian Sun, Fanglong Yao, Hongfeng Yu, and Li Jin are with the Key Laboratory of Network Information System Technology (NIST), Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China (e-mail: sunxian@aircas.ac.cn; yaofanglong17@mails.ucas.ac.cn; yuhf@aircas.ac.cn; jinlimails@gmail.com).

Jianhua Tao is with the Department of Automation, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100190, China (e-mail: jhtao@tsinghua.edu.cn).

Zhao Lv and Cunhang Fan are with the Anhui Province Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei, Anhui 230601, China (e-mail: kjlz@ahu.edu.cn; cunhang.fan@ahu.edu.cn).

Digital Object Identifier 10.1109/TPAMI.2024.3447778

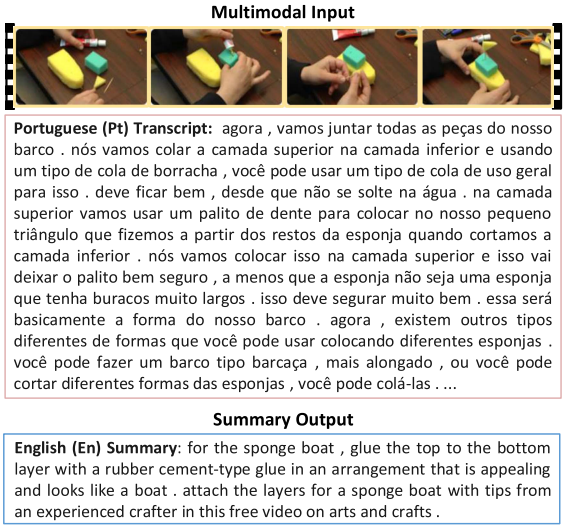

Fig. 1. An example of the MCLS task, which seeks to generate a target language (e.g., English) summary based on the video and its text in the source language (e.g., Portuguese).

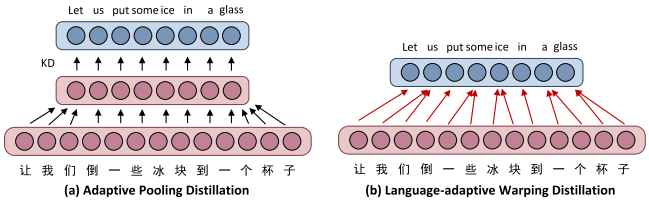

Fig. 2. Comparison between the adaptive pooling distillation and language-adaptive warping distillation (LAWD). Each coloured circle corresponds to the encoded feature vector of each word.

(CLS), a major challenge is that the limited number of bilingual experts makes the construction of a high-quality MCLS dataset resource-intensive. The experiment results of directly training on such limited multimodal data are typically under satisfactory, in which the low-resource language in the cross-lingual data limits the model to learn the language and video knowledge with relatively abundant resources.

To overcome the above issues in MCLS, we propose a knowledge distillation (KD) induced triple-stage training method, which leverages the knowledge of the MS model trained with prevalent monolanguage to assist the resource-constrained MCLS generation. In the first stage, we propose the video-guided dual fusion network (VDF) by designing dual diverse fusion strategies at both encoder and decoder structures to integrate multimodal and cross-lingual information. It is trained on sufficient multimodal summarization data in the target language. In the second stage, encoder- and vocab-level adaptive pooling distillation methods are proposed to adjust the output features between the target language encoder (vocab) in VDF model and a new source language encoder (vocab). In the third stage, the target encoder (vocab) is replaced with the KD-induced source encoder (vocab), composing a new VDF model. The new VDF is fine-tuned on the limited MCLS data and serves as the final model to generate the target language summaries.

Further, drawing inspiration from the challenge of variable-length pronunciation alignment in speech recognition [8], [9], we propose language adaptive warping distillation (LAWD), an improved alternative to adaptive pooling distillation in the above triple-stage training method. Fig. 2 provides an illustrative example of the two ways of knowledge transfer in a Chinese-English scenario. Adaptive pooling distillation forces language feature sequences of different lengths into the same length and then aligns them. Although it can bridge Chinese and English to a certain extent, the destruction of the language feature structure may potentially lead to information loss in adapting triple-stage training models. In contrast, LAWD allows direct knowledge transfer between source and target language feature sequences while keeping the original language feature shape and position unchanged, avoiding the potential information loss due to structural damage.

To simulate the MCLS, we construct the How2-MCLS dataset by reorganizing the How2 dataset [10], where the How2 dataset is a large-scale multimodal understanding dataset of open-domain videos in Portuguese and English. We also meticulously annotated MCLS data for a Chinese-to-English scenario through crowdsourced translation to further validate the robustness of the proposed method across different languages. Experiment results show that the VDF model alone achieves competitive performance in MCLS, and the performance further improves by a large margin via the proposed KD-induced triple-stage training method. What's more, with only 3k samples, our triple-stage training method outperforms strong baselines trained with more than 10 k samples.

This work is a substantial extension of the version presented at the EMNLP conference [11]. In EMNLP, we explored methods for transferring knowledge from the resource-rich MS model to assist the resource-constrained MCLS model, where we introduced the MCLS task, VDF backbone network, and a triple-stage training method driven by adaptive pooling distillation. In this work, we further propose a new method for knowledge distillation of variable-length parallel cross-lingual sequences, named language adaptive warping distillation (LAWD), as an alternative to adaptive pooling distillation. As far as we know, rather than forcing parallel language feature sequences of different lengths into the same shape through averaging, pooling, padding, or truncation, as in previous work, LAWD directly performs knowledge transfer while maintaining the shapes and positions of variable-length cross-lingual feature sequences, which avoids potential information loss caused by the disruption of language feature structures during knowledge distillation. Additionally, we expand the originally constructed How2-MCLS dataset and introduce a new Chinese-to-English MCLS dataset, and more experiments are conducted to analyze the effectiveness of our proposed method. We will release the newly added code and data[1].

In summary, the contribution of this work could be summarized as follows:

[1]The new code and data will be released at https://github.com/korokes/MCLS

1) We introduce multimodal cross-lingual summarization for videos (MCLS), to assist non-native viewers in understanding video contents in other languages. A video-guided dual fusion network (VDF) is proposed to make it applicable in MCLS.
2) We propose a KD-induced triple-stage training method to alleviate the problem of limited resources in MCLS, where encoder-level and vocab-level adaptive pooling distillation are designed to drive a VDF model that benefits from sufficient MS data in the prevalent language.
3) We propose a novel language-adaptive warping distillation (LAWD) applied in the triple-stage training method as an improved alternative to adaptive pooling distillation, which allows knowledge transfer of cross-lingual feature sequences of unequal length while keeping the original language structure unchanged.
4) We meticulously annotated the How2-MCLS dataset based on the How2 dataset to simulate the MCLS scenario. Experiment results illustrate that the proposed methods obviously outperform strong baselines, and can bring a lot of performance improvement to the MCLS model by migrating MS model knowledge.

II. Related Work

A. Multimodal Summarization

Unlike text summarization that only operates on the text modality, the objective of multimodal summarization (MS) [12], [13], [14], [15] is to compress multimodal data, such as text, audio, images, and videos, into a summary that reflects the salient information of the multimodal document. Multimodal summarization can be roughly divided into extractive methods [16] and abstractive methods [17]. Multimodal extractive summarization selects the most informative part of the multi-source data itself as the summary without generating content not included in the original input. Besides, multimodal abstractive summarization aims to generate abstractive summaries not present in the input document. In this work, we mainly focus on abstractive multimodal summarization.

In the early days, abstractive multimodal summarization was mainly focused on image and text modalities [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30]. For example, Chen et al. [18] and Li et al. [19] proposed a multimodal attention-based RNN encoder-decoder model that generates concise text for a set of news documents containing images. Li et al. [20] introduced a multimodal pointer network for generating text summaries of e-commerce product data. Benefiting from the success of pretrained language models, Lu et al. [21], Im et al. [22] and Lin et al. [24] integrated the pretrained BART [25] into multimodal summarization models for initializing the parameters of text encoders and decoders. Considering that existing abstractive text-and-image summarization pays little attention to visual quality, Liang et al. [26] proposed two summary-oriented visual auxiliary training strategies: image-to-summary generation in the absence of source text and masked image prediction [27]. Recently there are also several works [28] about MCLS. However, it's proposed for image modality rather than video modality.

Video-contained multimodal summarization generates research interest as the field advances [2], [5], [6], [7], [10], [14], [31], [32], [33], [34], [35]. Sanabria et al. [10] introduced a large-scale multimodal abstractive summarization dataset for open-domain videos, namely the How2 dataset, which provides various information sources including video, audio, textual transcripts, and manually generated summaries. Palaskar et al. [2] proposed a hierarchical attention based multi-source sequence-to-sequence model to integrate video and transcribed textual inputs to generate textual summaries. Liu et al. [3] proposed a multi-stage fusion network with forget gate to suppress cross-modal noise flow during multimodal fusion, and evaluated the models on both multimodal RNN and Transformer architectures. Yu et al. [5] explored how to incorporate video information into pre-trained language models to leverage their powerful generation capabilities for multimodal summarization. Considering that many works only utilize video and text modalities, or convert audio to text to utilize audio modality, some works [14], [32], [33] proposed to absorb audio acoustic features to facilitate generation of summaries. In multimodal text-image summarization, Zhu et al. proposed MSMO [17], which introduces the concept of generating multi-modal summaries in an abstractive approach. It selects visually salient images as visual summaries by internally assigning attention weights to the image set during the generation of textual summaries by the model. In video scenarios, VMSMO [6], XMSMO [7] were subsequently proposed, generating concise textual summaries for video news documents while selecting a video cover as an additional visual summary based on the matching degree of video and textual features [35]. However, those multimodal summarization methods mainly focus on monolingual scenarios, not on cross-lingual scenarios. In this work, we proposed adaptive pooling distillation and Language-Adaptive Warping Distillation (LAWD) to narrow this gap between different languages.

B. Cross-Lingual Summarization

In order to facilitate the ability to understand the content of articles in different languages, the task of cross-lingual summarization (CLS) [36] has been proposed, which aims to generate target language summaries for source language documents. Cross-lingual summarization can be broadly categorized into pipeline and end-to-end methods.

The pipeline-style cross-lingual summarization aims to divide the task into two stages: summarization and translation. Some work [37], [38], [39], [40], [41], [42] proposed the translate-then-summarize or the summarize-then-translate approach in cross-lingual summarization. For example, Wan et al. [38] proposed a graph-based sorting algorithm to generate cross-lingual summaries in a translate-then-summarize style, which uses the similarity of bilingual features of source documents and translated documents to extract summaries from translated documents. Considering that the summarize-then-translate approach can avoid the additional computational cost incurred by translating the entire document, Wan et al. [43] used an

extractive summarization method to generate candidate summary sentences in English and then employed a support vector machine model to predict the quality of each candidate sentence.

Despite the progress of pipeline methods, they suffer from error propagation and latency problem between translation model and summarization models [36], [44], and model combination complexity for summarization and translation models is higher in multilingual scenarios. On the other hand, end-to-end cross-lingual summarization models have gained researchers– interest [45], [46], [47], [48]. For example, Takase et al. [49] introduced multi-task learning into the encoder-decoder method for cross-lingual summarization, where the model was trained with single-language summarization, cross-lingual summarization, and machine translation tasks simultaneously. Liang et al. [50] proposed a variational hierarchical model for cross-lingual summarization. Xu et al. [51] employed four tasks, including masked language modeling, denoising autoencoder, monolingual summarization, and translation, to jointly pre-train the model and enhance its ability for cross-lingual summarization. Wang et al. [52] explored the performance of LLMs [53] on cross-lingual summarization and found that it prefers to produce lengthy summaries with detailed information. While, these methods do not cater to the needs of multimodal video scenarios. In addition, multimodal cross-lingual summarization task is likely to meet the data-insufficient problem. In this work, we propose the triple-stage training method to address the above problems, which aims to transfer knowledge from the resource-rich MS model to enhance the MCLS model.

III. Method

A. Overview

Given a video and its corresponding transcript text in the source language (e.g., Portuguese) as inputs, the goal of the MCLS system is to generate an abstractive summary in the target language (e.g., English). To alleviate the problem of limited resources in MCLS, it could be accompanied by English video-summary data and some parallel transcripts. Please note that there are multiple languages such as Portuguese and Chinese could be served as the source language. In this paper, we take Portuguese as the example source language to ensure clear and fluent expression.

Formally, let the video representations be $V = (v_1, \ldots, v_k)$, where v is the feature vector extracted by a pre-trained model. Besides, the English transcript and Portuguese transcript are denoted as $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_m)$, which consist of n and m tokens, respectively. The English summary could be denoted as a sequence of word tokens $S = (s_1, \ldots, s_l)$ consisting of several sentences. The task aims to predict the best summary sequence S by finding:

$$\arg\max_{\theta} Prob(S|X, Y, V; \theta) \tag{1}$$

where θ is the set of trainable parameters. If there is no assistance of the English transcript, we set $X = \{None\}$.

In the following sections, we first show the details of the VDF model for MCLS, and then describe our triple-stage training method that utilizes adaptive pooling distillation and language-adaptive warping distillation (LAWD) to drive a VDF model to assist MCLS models, where encoder-level and vocab-level knowledge distillation patterns are designed.

B. Video-Guided Dual Fusion Network

Fig. 3(c) illustrates the structure of the video-guided dual fusion network (VDF). VDF utilizes the language modality as the primary modality and the video as the guide modality to progressively integrate multimodal information. Dual fusion strategies are designed for VDF according to the source and target text characteristics during the encoding and decoding stages. We explain the VDF model using the given video, Portuguese transcript and English summary as an example.

1) Video and Text Encoder: Encoding Video. The video encoding features $V = (v_1, \ldots, v_k)$ are extracted from every 16 nonoverlapping frames by a pretrained action recognition model: a ResNeXt-101 3D convolutional neural network [54] trained for recognizing 400 different human actions in the Kinetics dataset [55].

$$V = \text{3DCNN}_{\text{ResNeXt-101}}(Frames) \tag{2}$$

We add learnable position embeddings for video features.

Encoding Transcript. We apply a standard bidirectional transformer encoder [56] to get contextual text encoding features. Take Portuguese transcript Y as an example, the encoding process could be denoted by the following equation:

$$T^Y_{Trm} = \text{BiTrm}(Y) \tag{3}$$

where "BiTrm" means the standard bidirectional transformer encoder layers. After completing the independent encoding of text and video, multimodal fusion is performed through the forget fusion encoder module.

2) Forget Fusion Encoder: Considering the source transcript text and video are lengthy and have much redundancy, forget fusion encoder first adopts the forget gate fusion (FGF) [3] module, which fuses video information to text while suppressing the flow of cross-modal noise. As illustrated in Fig. 4(a), assuming the input source transcript Y is in Portuguese, FGF incorporates relevant video information V into the text features T^Y_{Trm} and obtains a multimodal text representation T^Y_{FGF} with relevant video information:

$$T^Y_{FGF} = \text{FGF}(T^Y_{Trm}, V) \tag{4}$$

Then, the fusion feature T^Y_{FGF} is fed into a self-attention layer and a feed-forward layer to reconstruct its representation, obtaining the encoder output T^Y_{en}.

3) Cascaded Fusion Decoder: We propose a cascade fusion paradigm that progressively integrates multimodal and cross-lingual features to generate context vectors for decoding. It makes a fusion of the target language features and the attentive source language contextual features, and then integrates the fused language features and the attentive video contextual features. As illustrated in Fig. 4(b), the decoder consists of three sets of cascaded scaled dot-product attention layers: First, the target English summary features are fed to a masked

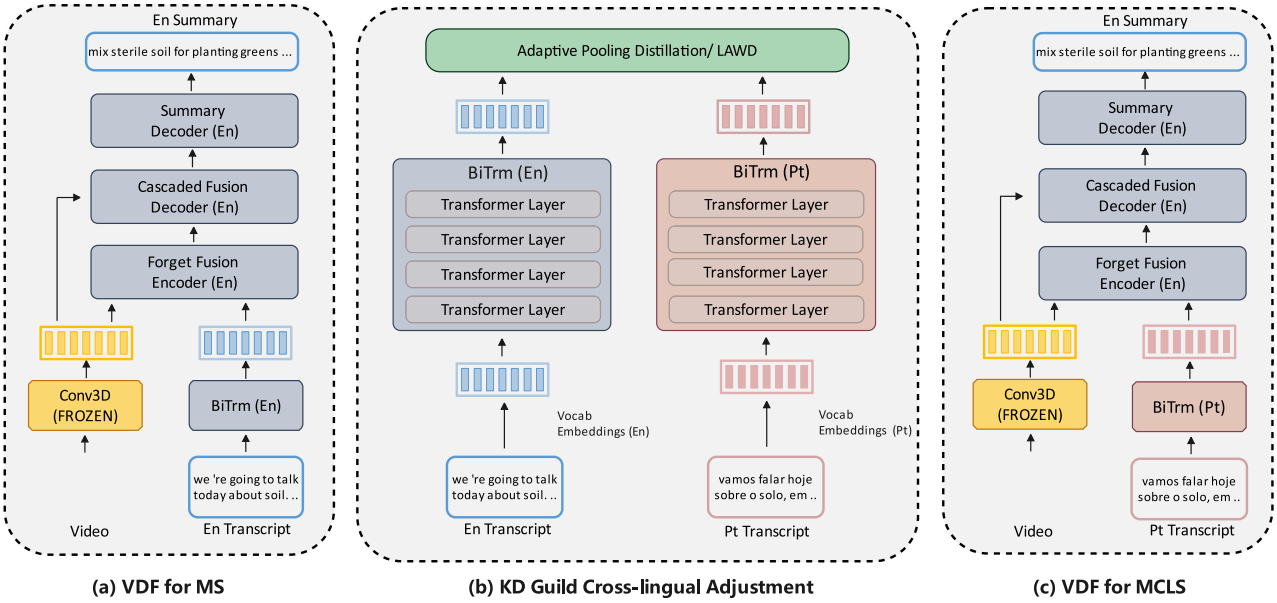

Fig. 3. The proposed triple-stage training method for MCLS in Video+Pt2En situation, which includes: (1) training a VDF model for multimodal summarization (MS) on English video-summary corpus; (2) cross-lingual encoder adjustment via knowledge distillation (KD); (3) replacing the English teacher encoder with the Portuguese student one, and generating English summaries with Portuguese videos as the input. Please note that in this work, we propose encoder/vocab-level KD in the second stage, and we only illustrate the former one for brevity.

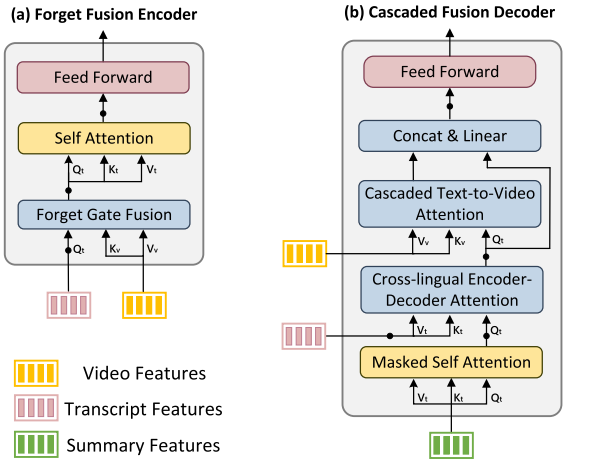

Fig. 4. The multimodal feature fusion strategies at encoding and decoding in VDF, include (a) forget fusion encoder and (b) cascaded fusion decoder, respectively. "·" represents a layer normalization operation.

self-attention layer, obtaining its context vector C^X. Then, the source Portuguese context vector C^Y is calculated through the cross-lingual encoder-decoder attention of the target English summary context C^X to source text encodings T^Y_{en}:

$$C^Y = \text{Att}(C^X, T^Y_{en}, T^Y_{en}) \tag{5}$$

Next, the Portuguese text context vector C^Y and the English summary context vector C^X are fused via a residual connection. The fused text context is used to calculate its attentive video context C^V via a cascaded text-to-video attention layer:

$$C^V = \text{Att}(C^X + C^Y, V, V) \tag{6}$$

Finally, C^X, C^Y, C^V are merged by a fusion layer to obtain the final multimodal context C^M by passing in the subsequent decoding structures. The text and video contexts are concatenated and fed into a linear layer with a residual connection to deepen the memory of original text information:

$$C^M = [(C^X + C^Y), C^V]W_{de} + b_{de} + (C^X + C^Y) \tag{7}$$

where W_{de}, b_{de} are learnable parameters. [..,..] is the concatenation operation.

C. Knowledge Distillation Induced Triple-Stage Training Method

To alleviate the problem of limited resources in MCLS, we further propose a triple-stage training method, which transfers the knowledge of the model trained with sufficient English MS data to the model under MCLS data. As illustrated in Fig. 3, it consists of the following three stages:

1) Stage 1, shown in Fig. 3(a), multimodal summarization for sufficient mono language videos: leveraging VDF described in III.B as the backbone, and training a model $\phi_{\mathcal{X}}$ that receives videos and English transcripts to generate English summaries.
2) Stage 2, shown in Fig. 3(b), knowledge distillation (KD) from the prevalent language to the objective one through parallel bilingual transcript data. We design two strategies: a) adaptive pooling distillation, which provides feature pooling operations to process the output feature sequences of the teacher and student models into the same length for cross-language adjustment; b) language-adaptive warping distillation (LAWD), which provides feature distortion operations on different languages without changing their

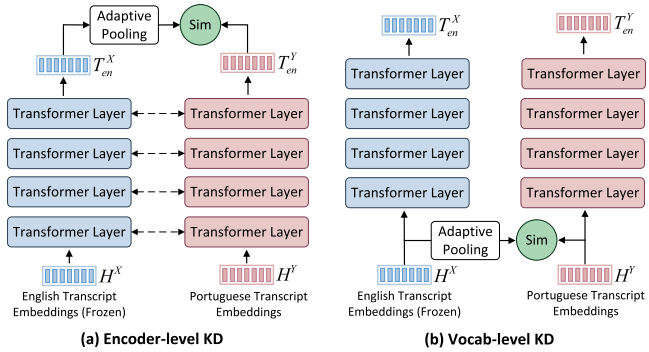

Fig. 5. Illustration of the adaptive pooling distillation, including: (a) encoder-level KD; (b) vocab-level KD.

shape and position for adjustment. Both the two KD methods are designed at encoder level and vocab level.

3) Stage 3, shown in Fig. 3(c), multimodal summarization for cross-lingual videos: replacing the English encoder $\phi_{E_\mathcal{X}}$ (vocab $\phi_{V_\mathcal{X}}$) in stage 1 with the Portuguese encoder $\phi_{E_\mathcal{Y}}$ (vocab $\phi_{V_\mathcal{Y}}$) in stage 2 to form a new VDF model $\phi_\mathcal{Y}$. Then fine-tune $\phi_\mathcal{Y}$ on Portuguese transcripts and videos to generate English summaries.

D. Cross-Lingual Adjustment Via Knowledge Distillation

In this section, we introduce the knowledge distillation (KD) mechanism to leverage the prior knowledge in English and assist the summary generation. It utilizes a teacher-student method to transfer the knowledge from the target language (teacher) to the source language (student). Assume we have trained a well-informed VDF model $\phi_\mathcal{X}$ on the English multimodal summarization data. By looking up the vocabulary embedding table $\phi_{V_\mathcal{X}}$, the English transcript X are converted to the embedding matrix $H^X \in \mathbb{R}^{n\times e}$, where n and e are the input English sequence length and the dimension of the embeddings, respectively. After feeding H^X to the encoder $\phi_{E_\mathcal{X}}$ of VDF $\phi_\mathcal{X}$, we could obtain the English encoder output $T^X_{Trm} \in \mathbb{R}^{n\times e}$.

During the KD process, a crucial problem is that the parallel sequence lengths of different language transcripts are diverse. To overcome this problem, we offer two KD mechanisms (i.e., adaptive pooling distillation and LAWD). For each mechanism, encoder-level KD and vocab-level KD are designed.

1) Adaptive Pooling Distillation: Based on the components of VDF, there are two options to transfer the prior knowledge from English to Portuguese:

1) Encoder-level KD: as illustrated in Fig. 5(a), a new randomly initialized Portuguese encoder $\phi_{E_\mathcal{Y}}$ is trained from scratch, which has the same architecture as the English encoder $\phi_{E_\mathcal{X}}$. We could obtain the Portuguese encoder output T^Y_{Trm} after feeding the Portuguese transcripts Y. During the encoder-level KD process, we treat the English encoder $\phi_{E_\mathcal{X}}$ as the teacher model and the Portuguese encoder $\phi_{E_\mathcal{Y}}$ as the student model. The goal of encoder-level KD is to transfer the knowledge from $\phi_{E_\mathcal{X}}$ to $\phi_{E_\mathcal{Y}}$ by making the output feature distributions consistent for different languages.

Algorithm 1: Encoder-Level Adaptive Pooling Distillation.

Input: target language token sequence X, source language token sequence Y, frozen teacher model $\phi_{E_\mathcal{X}}$, learnable student model $\phi_{E_\mathcal{Y}}$
Output: learnable student model $\phi_{E_\mathcal{Y}}$ (finish)
$T^X_{Trm} = \phi_{E_\mathcal{X}}(X) \in \mathbb{R}^{n\times e}$ // target language feature
$T^Y_{Trm} = \phi_{E_\mathcal{Y}}(Y) \in \mathbb{R}^{m\times e}$ // source language feature
for $i = 1, \ldots, m$ **do** // adaptive pooling strategy
  $lstart = \text{floor}(i \times n \div m)$
  $lend = \text{ceil}((i+1) \times n \div m)$
  $T^{X\to Y}_{Trm}[i] = \frac{\text{sum}(T^X_{Trm}[lstart{:}lend])}{lstart - lend}$
end for
$\mathcal{L}_{kd} = \text{smoothL1}(T^{X\to Y}_{Trm}, T^Y_{Trm})$ // loss of distillation
$\phi_{E_\mathcal{Y}}[parm] := \phi_{E_\mathcal{Y}}[parm] - \alpha \frac{\partial \mathcal{L}_{kd}}{\partial E^Y_{Trm}}$ // optimize the student model via loss backpropagation

Considering that the language features $T^X_{Trm} \in \mathbb{R}^{n\times e}$ and $T^Y_{Trm} \in \mathbb{R}^{m\times e}$ output by the teacher and student models have different matrix structures and cannot be directly distilled, we introduce an adaptive pooling strategy to transform the English feature sequence to the same length as the Portuguese one. Algorithm 1 gives the pseudocode. After that, The goal is to align source language feature T^Y_{Trm} and target language feature $T^{X\to Y}_{Trm} \in \mathbb{R}^{m\times e}$ that changes the shape. Smooth L1 loss is leveraged to optimize the encoder-level KD.

$$\mathcal{L}_{kd} = \text{smoothL1}(T^{X\to Y}_{Trm}, T^Y_{Trm}) \tag{8}$$

2) Vocab-level KD: as shown in Fig. 5 (b), the vocab-level KD process directly operates on the vocab embeddings. Specifically, we look up a learnable randomly initialized Portuguese vocab embedding table $\phi_{V_\mathcal{Y}}$ and receive the Portuguese embedding H^Y. We fix the parameters of the English vocab embedding table and transfer its knowledge to the Portuguese vocab embedding table. Similarly, the embedding feature H^X output by the teacher model $\phi_{V_\mathcal{X}}$ is transformed into the same shape ($H^{X\to Y}$) as the embedding feature H^Y output by the student model $\phi_{V_\mathcal{Y}}$ through the adaptive pooling strategy. Then the vocab-level KD process is fulfilled by minimizing the smooth L1 loss between the English embedding $H^{X\to Y}$ and the Portuguese embedding H^Y.

$$\mathcal{L}_{kd} = \text{smoothL1}(H^{X\to Y}, H^Y) \tag{9}$$

2) Language-Adaptive Warping Distillation: In the process of adaptive pooling distillation, the teacher's language feature shape changes to match the student, resulting in a certain degree of information loss. More importantly, the alteration in input shape after pooling is incompatible with model trained in first-stage, which only fits the original inputs without pooling operation. As a result, we introduce the language-adaptive warping distillation (LAWD), as illustrated in Fig. 6, which enables the different length features output from the teacher and student

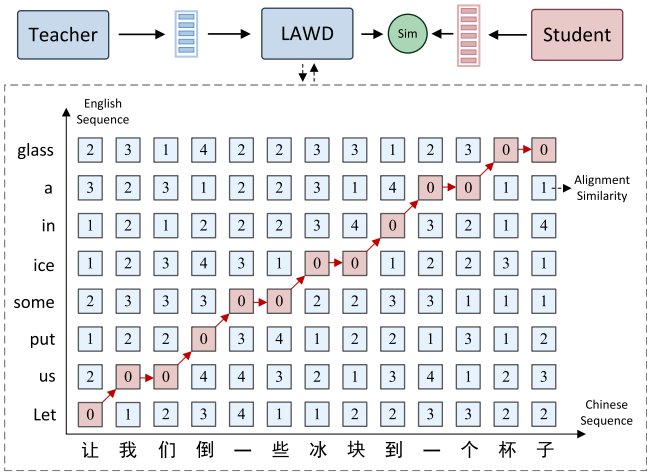

Fig. 6. Illustration of the proposed LAWD, where the numbers inside the boxes represent alignment similarities. We randomly set the values of the alignment similarities inside each box for demonstration purposes. The numbers in actual situation may be different.

Algorithm 2: Vocab-Level Language-Adaptive Wraping Distillation (LAWD).

Input: target language token sequence X, source language token sequence Y, frozen teacher model ϕ_{V_X}, learnable student model ϕ_{V_Y}
Output: learnable student model ϕ_{V_Y} (finish)
$H^X = \phi_{V_X}(X) = (h_1^X, \ldots, h_n^X) \in \mathbb{R}^{n\times e}$ // target language feature
$H^Y = \phi_{V_Y}(Y) = (h_1^Y, \ldots, h_m^Y) \in \mathbb{R}^{m\times e}$ // source language feature
$C(H^X, H^Y) = (c_{i,j})_{n\times m} = (\|h_i^X - h_j^Y\|)_{n\times m} \in \mathbb{R}^{n\times m}$ // teacher-student cost matrix
for $j = 1, \ldots, m$ **do**
  for $i = 1, \ldots, n$ **do**
    $r_{i,j} = -\log(e^{-r_{i-1,j-1}} + e^{-r_{i-1,j}} + e^{-r_{i,j-1}}) + c_{i,j}$
  end for
end for
$\mathcal{L}_{kd} = r_{n,m}$ // loss of distillation
$\phi_{V_Y}[parm] := \phi_{V_Y}[parm] - \alpha\frac{\partial \mathcal{L}_{kd}}{\partial H^Y}$ // optimize the student model via loss backpropagation

models to be distilled while keeping the original language feature structure unchanged.

In this section, we take the vocab-level LAWD as an example. Please note that LAWD can also be applied for the encoder-level distillation scenario. The pseudocode of the vocab-level LAWD is shown in Algorithm 2.

Specifically, given target language embedding matrix $H^X = (h_1^X, \ldots, h_n^X) \in \mathbb{R}^{n\times e}$ from frozen teacher model and source language embedding matrix $H^Y = (h_1^Y, \ldots, h_m^Y) \in \mathbb{R}^{m\times e}$ from learnable student model, we define a teacher-student cost matrix $C(H^X, H^Y) = (c_{i,j})_{n\times m} = (\|h_i^X - h_j^Y\|)_{n\times m} \in \mathbb{R}^{n\times m}$ that gauges all pairwise affinities, where euclidean distance $\|\cdot\|$ is used to measure the distance. For example, $c_{i,j} = \|h_i^X - h_j^Y\|$ denotes the euclidean distance from the i-th source language token vector to the j-th target language token vector. The goal of LAWD aims to find and optimize the path cost from coordinates $(0,0)$ to (n,m) to reduce the distance between student features and teacher features, in which three path directions of right, upper, and upper right are retained to reduce the search space. For example, the red grid shown in Fig. 6 forms a path. Therefore, in this process, the characteristics of teachers and students can protect the original structure from being damaged.

Based on the above description, $R = (r_{i,j})_{n\times m} \in \mathbb{R}^{n\times m}$ is defined as the minimum cost path matrix derived from the teacher-student cost matrix C, where $r_{i,j}$ represents the minimum cost of path from the source language sequence $H^Y[0:j]$ to the target language sequence $H^X[0:i]$. Dynamic programming is used to solve the minimum cost of path from $(0,0)$ to (n,m) coordinates:

$$r_{i,j} = \min\{r_{i-1,j-1}, r_{i-1,j}, r_{i,j-1}\} + c_{i,j} \tag{10}$$

where $r_{i-1,j-1}, r_{i-1,j}, r_{i,j-1}$ respectively represent the three directions of advancement in the path: upper right, upper, and right. To make the distillation process learnable, the discrete minimum calculation min(·) are replaced by a differentiable continuous calculation, defined as:

$$S(r_{i-1,j-1}, r_{i-1,j}, r_{i,j-1}) \triangleq -\log(e^{-r_{i-1,j-1}} + e^{-r_{i-1,j}} + e^{-r_{i,j-1}}) \tag{11}$$

So the (10) becomes:

$$r_{i,j} = S(r_{i-1,j-1}, r_{i-1,j}, r_{i,j-1}) + c_{i,j} \tag{12}$$

Actually, it changes from a discrete optimal path cost ((10)) to an optimizable mapping of the sum of all path costs based on three preset forward directions. So, for the feature sequence H^X of length m from the frozen teacher model and the feature sequence H^Y of length n from the learnable student model, the loss of distillation could be:

$$\mathcal{L}_{kd} = r_{n,m} \tag{13}$$

which aims to minimize the cost of all path combinations from $(0,0)$ to (n,m) coordinates in the teacher-student cost matrix $C(H^X, H^Y)$.

In loss backpropagation, based on the chain rule, we could obtain:

$$\left.\frac{\partial \mathcal{L}_{kd}}{\partial H^Y}\right|_{\mathcal{L}_{kd}=r_{n,m}} = \left(\frac{\partial C(H^X, H^Y)}{\partial H^Y}\right)^T \frac{\partial r_{n,m}}{\partial C(H^X, H^Y)} \tag{14}$$

where the term $(\frac{\partial C(H^X,H^Y)}{\partial H^Y})$ is the Jacobian matrix of C with respect to H^Y. For the term $\frac{\partial r_{n,m}}{\partial C(H^X,H^Y)}$, we could obtain:

$$\begin{aligned}\frac{\partial r_{n,m}}{\partial C} &= \left(\frac{\partial r_{n,m}}{\partial c_{i,j}}\right)_{n\times m} = \left(\frac{\partial r_{n,m}}{\partial r_{i,j}}\frac{\partial r_{i,j}}{\partial c_{i,j}}\right)_{n\times m} \\ &= \left(\frac{\partial r_{n,m}}{\partial r_{i,j}}\frac{\partial (c_{i,j} + S(r_{i-1,j-1}, r_{i-1,j}, r_{i,j-1}))}{\partial c_{i,j}}\right)_{n\times m} \\ &= \left(\frac{\partial r_{n,m}}{\partial r_{i,j}}\right)_{n\times m}\end{aligned} \tag{15}$$

for the term $(\frac{\partial r_{n,m}}{\partial r_{i,j}})$ in (15), the time complexity of using automatic derivation[2] is $O(m^2n^2)$. The time complexity of backpropagation could be optimized based on path search space constraints. Based on the chain rule, there are:

$$\frac{\partial r_{n,m}}{\partial r_{i,j}} = \frac{\partial r_{n,m}}{\partial r_{i+1,j}}\frac{\partial r_{i+1,j}}{\partial r_{i,j}} + \frac{\partial r_{n,m}}{\partial r_{i,j+1}}\frac{\partial r_{i,j+1}}{\partial r_{i,j}} + \frac{\partial r_{n,m}}{\partial r_{i+1,j+1}}\frac{\partial r_{i+1,j+1}}{\partial r_{i,j}} \tag{16}$$

that is, only three preset directions for path advancement are retained in the chain rule calculations. Among them, for the terms $\frac{\partial r_{i+1,j}}{\partial r_{i,j}}$, $\frac{\partial r_{i,j+1}}{\partial r_{i,j}}$, and $\frac{\partial r_{i+1,j+1}}{\partial r_{i,j}}$, we take $\frac{\partial r_{i+1,j}}{\partial r_{i,j}}$ as an example to perform the calculation:

$$\frac{\partial r_{i+1,j}}{\partial r_{i,j}} = \frac{\partial\left(-\log(e^{-r_{i,j-1}} + e^{-r_{i,j}} + e^{-r_{i+1,j-1}}) + c_{i+1,j}\right)}{\partial r_{i,j}} = \frac{e^{-r_{i,j}}}{e^{-r_{i+1,j-1}} + e^{-r_{i,j}} + e^{-r_{i,j-1}}} \tag{17}$$

Taking the logarithm of both sides of 17, there are:

$$\log\frac{\partial r_{i+1,j}}{\partial r_{i,j}} = S\left(r_{i+1,j-1}, r_{i,j}, r_{i,j-1}\right) - r_{i,j} = r_{i+1,j} - c_{i+1,j} - r_{i,j} \tag{18}$$

Likewise, we could obtain:

$$\frac{\partial r_{i+1,j}}{\partial r_{i+1,j}} = e^{r_{i+1,j} - c_{i+1,j} - r_{i,j-1}} \tag{19}$$

$$\frac{\partial r_{i,j+1}}{\partial r_{i,j}} = e^{r_{i,j+1} - c_{i,j+1} - r_{i,j}} \tag{20}$$

$$\frac{\partial r_{i+1,j+1}}{\partial r_{i+1,j}} = e^{r_{i+1,j+1} - c_{i+1,j+1} - r_{i,j}} \tag{21}$$

For the terms $\frac{\partial r_{n,m}}{\partial r_{i+1,j}}$, $\frac{\partial r_{n,m}}{\partial r_{i,j+1}}$, and $\frac{\partial r_{n,m}}{\partial r_{i+1,j+1}}$ in (16), the results from $\frac{\partial r_{n,m}}{\partial r_{n,m}} = 1$ to $\frac{\partial r_{n,m}}{\partial r_{1,1}}$ can be deduced. Then the time complexity of (16) is reduced to $O(mn)$ to improve the efficiency of loss backpropagation.

E. Multimodal Cross-Lingual Summarization for Videos

After the KD process, the Portuguese output distribution and the real English distribution could be as similar as possible. As a result, the remaining fractions of VDF that have not been distilled can deal with cross-lingual multimodal summarization just like with monolingual summarization. On the basis of this, as illustrated in Fig. 3(c), we could replace the English encoder ϕ_{D_x} (vocab ϕ_{V_x}) with the Portuguese encoder ϕ_{E_y} (vocab ϕ_{V_y}) to form a new VDF model ϕ_y. Finally, we utilize the new VDF model ϕ_y to achieve the Portuguese video V and transcript Y, and fine-tune it with cross-entropy loss by calculating the output summary word probability $\hat{S}$ between the gold summary S:

$$\hat{S} = \phi_y(V, Y) \tag{22}$$

[2]torch.Tensor.backward()

Process 3: Annotation of How2-MCLS Dataset in Video+ Zh2En MCLS Scenario.

Input: Raw English corpus following MMT partitioning in How2 dataset
Output: Translated Chinese corpus applicable to MCLS

```
for 4 iterations do
    for video transcript text in corpus do
        for sentence in video transcript text do
            translate sentence via translation API
            post-editing
        end for
    end for
    correct full-transcript consistency
end for
randomly inspect 10% of annotation results
if pass then
    break
end if
end for
```

$$\mathcal{L}_{ft} = -\sum_{t=l}^{L}\log P\left(S_t \mid \hat{S}_{<t}, V, Y\right) \tag{23}$$

IV. Experiments

A. Dataset Construction

We conduct the experiments by reorganizing the How2 dataset [10]. How2 dataset is a large-scale open-domain instructional video dataset, which includes three sub-task data: multimodal machine translation (MMT) [57], [58], [59], multimodal speech recognition (MSR) [60], [61], and multimodal summarization (MS). The MS data includes 2,000 h of videos accompanied by English transcripts and summaries. Besides, MMT data includes 300 h of videos combined with bilingual Portuguese and English transcripts. Between the two, the data in MS contains those from MMT.

For simulating MCLS, we utilize MMT's 300 h videos and bilingual transcripts, combined with the summaries provided by MS, as **How2-MCLS** dataset; and we refer to the official division approach from MMT to split the dataset for training, validation, and testing. In addition, we exclude data in MS appearing in How2-MCLS dataset and use the remaining English video-summary data of MS as support data to form **How2-MS** dataset.

Furthermore, we expanded the originally constructed How2-MCLS dataset by adding the Zh-En cross-lingual pair. We employed a total of 15 undergraduate and graduate students from Anhui University to complete the annotation of the Chinese and English corpus. Each annotator is a native Chinese speaker and has passed the CET 6 exam or obtained a score of 6 in the IELTS exam, ensuring that they can convert the English transcript into fluent Chinese with the aid of machine translation tools. Process 3 illustrates the annotation workflow. Specifically, the dataset was evenly divided into 15 parts and assigned to the annotators. Each document was segmented into multiple sentences according to the format of the MMT task in the How2

TABLE I
STATISTICS OF THE DATA PARTITION, WHERE THE NUMBERS REPRESENT THE NUMBER OF VIDEOS

Partition	Training	Validation	Test
How2-MCLS: Video+Pt2En	13,167	150	127
How2-MCLS: Video+Zh2En	10,692	1,325	1,310
How2-MS: Video+En2En	59,539	-	-

The data in how2-mcls and how2-ms datasets do not overlap, and the videos in how 2-mcls also have corresponding bilingual transcripts.

TABLE II
STATISTICS OF ANNOTATED CHINESE CORPUS

	Transcript	Summary
Total number of texts	13,327	13,327
Average text length	374.5	57.6
Maximum text length	1,515	130
minimum text length	30	9
Total number of sentences	193,652	22,273
Average number of sentences	14.5	1.7
Maximum number of sentences	88	4
minimum number of sentences	1	1
Total number of words	4,990,904	767,180

dataset. Each sentence was first translated into Chinese using a machine translation API, and then the annotators performed post-editing on the machine translation results to correct errors in content, grammar, spelling, punctuation, format, and terminology, ensuring the translation was accurate and fluent. Subsequently, based on the video information and the entire document, each sentence was further modified to maintain paragraph-level coherence and consistency. For each completed annotated part, 10% were randomly selected for manual quality checks to assess the accuracy and fluency of the annotation results, and a rating of "no modification required" or "still needs modification" is made. If the proportion of "no modification" ratings of the inspected data reaches 95%, the quality check was considered passed. Otherwise, annotators had to iteratively edit and revise based on the check feedback. We completed the translation annotation through four iterations. The overall division of the dataset is shown in Table I. For the annotated Chinese corpus, the statistics related to documents, sentences, and words is shown in Table II.

B. Implementation Details and Evaluation Metrics

Our model adoptd 4-layer, 512-dimensional, 8-head transformer encoder layers and decoder layers, and 1-layer forget fusion encoder layer and cascaded fusion decoder layer. The maximum text sequence length and video sequence length are truncated to 800 and 1024, respectively. For training, the proposed models are trained with a batch size of 8, and we adopt the Adam optimizer with the initial learning rate of 1.5e-4. For prediction, we use the beam search with a beam size of 6 and a length penalty as 1. Following Palaskar et al. [2], the video features are extracted from a ResNeXt-101 3D convolutional neural network, and the dimension is 2048. The vocabulary is constructed based on the How2 dataset, and we do not use pre-trained word embeddings.

In terms of evaluation metrics, we mainly use ROUGE (1,2,L)[3] [62], which are commonly used evaluation metrics for summarization. BLEU (1,2,3,4) [63], METEOR [64] and CIDEr[4] [65] are also adopted to analyze the experimental results comprehensively, which are metrics used to evaluate text generation tasks [66].

C. Backbone Performance

We construct recent multimodal summarization methods of single or multiple modalities as backbones for comparison: **S2S** [67]: a standard sequence-to-sequence architecture with attention mechanism for abstractive summarization. **NCLS** [68]: a transformer-based encoder-decoder model for cross-lingual summarization. **VideoRNN** [2]: a sequence-to-sequence RNN model that receives video features to generate summaries. **MT** [69]: a transformer-based encoder-decoder architecture transforming video sequence features to captions. **HA** [2]: a multisource sequence-to-sequence model with a hierarchical attention to combine video and text modalities. **MFFG** [4]: a multistage fusion network with the forget gate module for multimodal summarization. **Pipe-[Model]** and **Pipe-[Model]**†: "Translate-then-summarize" and "summarize-then-translate" pipeline methods of cross-lingual summarization [36], respectively, where the former translates the source language document into the target language and then conduct monolingual summarization, and the latter first summarizes the source monolingual document and then translates it into the target language. We use mBART[5] [70] as the translation model and fine-tune mBART through the parallel bilingual data of How2-MCLS to strengthen pipeline baselines. **Pipe-[Model]**$^{\circ}$ and **Pipe-[Model]**$^{\circ\dagger}$: The same settings as Pipe-[Model] and Pipe-[Model]†, but replace the fine-tuned mBART with the Google-translation-v2 API as the translator. We evaluated the pipeline architecture of NCLS, HA and MFFG.

In Tables III and IV, we simulate the MCLS by only using the How2-MCLS dataset to train models without accessing bilingual transcripts or How2-MS data. Besides, in pipeline baselines, bilingual transcripts were used to fine-tune the mBART translation model. From the experiment results, we can observe that: 1) Multimodal models typically perform better than unimodal models, illustrating the importance of multi-source information to promote summary generation. 2) Translate-then-summarize models generally outperform summarize-then-translate models. One possible reason is that in translate-then-summarize models most of the original information is still retained for summarization, and the decoder is directly supervised by ground-truth summaries. 3) Pipe-[Model] (Pipe-[Model]†) generally outperforms Pipe-[Model]Â° (Pipe-[Model]$^{\circ\dagger}$). This is because the pre-trained translation model mBART has been fine-tuned with How2 translation data distributed the same as the task data, which enhances the translation ability in the task domain.

[3][Online], Available: https://github.com/neural-dialogue-metrics/rouge
[4][Online], Available: https://github.com/Maluuba/nlg-eval
[5][Online], Available: https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

TABLE III
BACKBONE PERFORMANCE OF DIFFERENT MODELS ON THE HOW2-MCLS DATASET (PT-EN)

Modality	Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	CIDEr
Pt transcript	S2S	41.85	27.97	21.19	17.80	43.21	20.48	37.02	17.65	0.709
	NCLS	42.15	29.49	22.97	18.44	43.44	21.93	37.48	18.23	0.867
	Pipe-NCLS°	37.77	25.54	19.12	14.75	40.76	18.54	34.84	16.41	0.582
	Pipe-NCLS	39.62	27.42	20.77	16.35	41.54	19.92	35.97	17.04	0.674
Video	VideoRNN	34.98	22.97	16.94	12.69	38.74	16.98	33.97	14.95	0.469
	MT	35.96	23.72	17.66	13.39	39.25	17.17	34.05	15.12	0.580
Pt transcript+Video	HA	42.29	29.46	22.75	18.28	44.16	22.18	38.33	17.98	0.871
	MFFG	42.75	30.53	23.87	19.35	45.21	23.22	40.11	18.81	0.937
	Pipe-HA	41.61	28.43	21.40	16.73	43.32	20.51	37.47	17.95	0.784
	Pipe-MFFG	41.26	28.62	22.10	17.75	44.66	21.82	38.65	18.03	0.863
	Pipe-HA°	41.35	28.54	21.83	17.39	43.45	20.66	36.82	17.56	0.747
	Pipe-MFFG°	42.46	29.79	23.05	18.47	43.86	21.62	38.24	18.47	0.768
	VDF	**43.81**	**31.23**	**24.89**	**20.18**	**46.06**	**24.37**	**40.50**	**19.22**	**1.064**

Bold values represent the best results for each metric.

TABLE IV
BACKBONE PERFORMANCE OF DIFFERENT MODELS ON THE HOW2-MCLS DATASET (ZH-EN)

Modality	Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	CIDEr
Zh transcript	S2S	39.10	26.67	20.09	15.44	42.29	19.86	36.27	17.09	0.563
	NCLS	40.83	28.09	21.25	16.45	42.81	20.49	37.07	17.62	0.593
	Pipe-NCLS	40.46	27.27	20.34	15.61	41.28	18.99	35.82	16.94	0.507
	Pipe-NCLS°	39.05	25.58	18.55	13.69	39.79	17.19	34.31	16.13	0.368
	Pipe-NCLS†	36.08	24.07	17.49	12.69	41.21	18.67	35.26	16.00	0.368
	Pipe-NCLS°†	35.87	22.45	15.58	10.45	39.74	16.06	30.79	15.81	0.339
Video	VideoRNN	35.89	23.73	17.48	12.93	38.80	17.14	33.89	14.98	0.430
	MT	37.36	24.91	18.46	13.79	39.60	17.87	34.53	15.53	0.455
Zh transcript+Video	HA	41.50	28.66	21.85	17.12	43.79	21.27	37.64	18.04	0.692
	MFFG	43.60	31.30	24.70	20.13	46.01	24.06	40.21	19.21	0.881
	Pipe-HA	41.33	27.98	20.92	16.10	42.41	19.56	36.28	17.43	0.574
	Pipe-MFFG	42.26	29.31	22.35	17.53	44.34	21.71	38.17	18.22	0.689
	Pipe-HA°	41.12	27.30	20.01	15.04	41.38	18.31	35.03	17.17	0.457
	Pipe-MFFG°	42.90	29.57	22.52	17.68	43.66	21.05	37.38	18.32	0.634
	Pipe-HA†	39.12	26.11	19.07	14.12	41.21	18.67	35.26	16.75	0.425
	Pipe-MFFG†	39.59	26.98	20.14	15.27	42.23	19.99	36.38	17.28	0.491
	Pipe-HA°†	38.97	23.94	16.46	11.16	40.45	15.82	31.02	16.59	0.383
	Pipe-MFFG°†	39.17	24.65	17.35	12.06	41.05	16.84	31.19	17.07	0.426
	VDF	**44.24**	**32.07**	**25.44**	**20.82**	**46.49**	**24.80**	**40.84**	**19.53**	**0.920**

Bold values represent the best results for each metric.

4) Despite using a powerful translator API or a fine-tuned mBART model, the pipeline methods don't perform as well as the end-to-end models with the same configuration. A crucial reason is the error propagation between translation and summarization stages. 5) The performance of the pure video modality model is modest, which is due to the use of frozen video features extracted from the action recognition backbone that cannot be trained. 6) As a backbone network, the proposed VDF outperforms the recent multimodal baselines, demonstrating the effectiveness of the proposed dual fusion strategies in VDF. Overall, the experimental results in Tables III and IV under different language scenarios express similar observations.

D. Overall Performance

We evaluate the proposed KD-induced triple-stage training method, which transfers knowledge from a multimodal monolingual summarization model trained on the How2-MS dataset. For the proposed methods, we evaluate four model variants: **VDF-TS-V(E)**: This utilizes a Triple-Stage training method with VDF as its backbone, incorporating Vocab (Encoder)-level adaptive pooling distillation. **VDF-TS-V(E)***: Building on VDF-TS-V and VDF-TS-E, which replaces adaptive pooling distillation with the LAWD strategy.

First, we formulated three comparison variants to leverage multimodal monolingual summarization data: a) **[Model] (P+F)**: first pre-training on How2-MS and then fine-tuning on How2-MCLS. b) **[Model] (M)**: mix How2-MS and How2-MCLS for training. c) **Pipe-[Model] (P)**: pipeline method (described in Section IV-C), using a fine-tuned mBART [25] or Google translator to translate the source language text into the target language, and then use the monolingual multimodal summarization model trained with How2-MS to generate the summary. We also constructed these variables on the recent MFFG for comparison. The experiment results are shown in

TABLE V
OVERALL PERFORMANCE OF DIFFERENT MODELS ON THE HOW2-MCLS DATASET (PT-EN)

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	CIDEr
MFFG (P+F)	46.49	35.74	29.63	25.46	49.13	28.60	43.78	21.39	1.451
MFFG (M)	40.73	29.24	23.38	19.67	42.73	23.15	38.62	17.78	0.976
VDF (P+F)	48.35	37.47	31.88	28.06	51.19	31.98	45.78	22.22	1.727
VDF (M)	41.51	30.60	24.77	20.74	43.83	24.25	39.59	18.36	1.070
Pipe-VDF (P)	48.92	38.29	32.26	28.17	51.55	32.15	46.65	22.47	1.681
Pipe-VDF (P)°	49.14	37.67	31.38	27.08	49.97	29.54	44.31	22.67	1.347
MFFG-TS-E	46.97	36.11	30.31	26.52	49.74	29.60	44.15	21.69	1.753
VDF-TS-E	50.01	39.26	33.47	29.50	52.16	33.31	47.18	23.19	1.910
VDF-TS-E*	50.83	40.69	35.16	31.36	52.63	34.15	48.03	24.17	2.005
MFFG-TS-V	47.69	36.62	30.68	26.65	50.70	30.50	45.47	22.16	1.726
VDF-TS-V	49.37	38.82	33.13	29.26	51.75	33.14	46.95	22.88	1.875
VDF-TS-V*	**51.92**	**41.51**	**35.75**	**31.82**	**54.02**	**35.49**	**48.90**	**24.45**	**2.065**

P: pre-training; F: fine-tuning; M: mix-training. [model]-ts-v(e) denotes triple-stage training method with [model] backbone and vocab(encoder)-level adaptive pooling distillation, and * means lawd is adopted to replace adaptive pooling distillation.

TABLE VI
OVERALL PERFORMANCE OF DIFFERENT MODELS ON THE HOW2-MCLS DATASET (ZH-EN)

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	CIDEr
MFFG (P+F)	48.83	38.34	32.50	28.40	51.76	32.55	46.85	22.88	1.590
MFFG (M)	33.61	21.95	15.71	11.05	36.87	15.94	32.39	13.81	0.285
VDF (P+F)	49.86	39.26	33.39	29.28	51.95	32.84	47.05	23.18	1.631
VDF (M)	36.75	25.17	18.99	14.62	38.54	18.42	33.98	15.00	0.458
Pipe-VDF (P)	50.19	39.08	32.89	28.58	51.63	31.63	46.31	23.13	1.486
Pipe-VDF (P)°	48.85	37.41	31.09	26.80	49.62	29.24	43.85	22.41	1.382
MFFG-TS-E	49.84	39.10	33.14	29.00	52.17	32.63	47.12	23.24	1.608
VDF-TS-E	50.69	39.93	33.96	29.79	52.34	33.01	47.33	23.48	1.614
VDF-TS-E*	50.89	40.25	34.32	30.18	52.70	33.51	47.76	23.58	1.675
MFFG-TS-V	50.08	39.57	33.71	29.62	52.37	33.31	47.52	23.49	1.636
VDF-TS-V	50.73	40.36	34.57	30.47	52.89	34.11	47.98	23.80	1.702
VDF-TS-V*	**51.29**	**41.06**	**35.33**	**31.31**	**54.07**	**35.39**	**49.03**	**24.21**	**1.851**

Bold values represent the best results for each metric.

Tables V and VI, where we could find: 1) The proposed methods, VDF-TS-V(E) and VDF-TS-V*(E*), achieve the state-of-the-art performance. In particular, VDF-TS-V* performed better than VDF-TS-V, and VDF-TS-E* performed better than VDF-TS-E, which demonstrates the superiority of LAWD over adaptive pooling distillation in avoiding information loss caused by input shape changes. Among them, VDF-TS-V* achieves the best performance, which is due to the maximum knowledge transfer of MS model and the minimal parameter cost of learning cross-lingual alignment via LAWD. 2) Under the same experimental configuration, models with VDF as the backbone consistently outperformed those based on the strong baseline MFFG, illustrating the effect of the proposed VDF model. 3) Among our comparison variants, [Model] (M) generally do not perform well. This experiment result suggests that with imbalanced multilingual data, the model tends to converge towards the more abundant domain, hampering learning for the scarce domain. In particular, this phenomenon is more obvious in the Zh2En scenario because of the disparity in language feature distribution between Chinese and English. 4) We also conduct a KD-induced triple-stage training method for the MFFG model, i.e., MFFG-TS-E and MFFG-TS-V. The results suggest that, with the same backbone, MFFG-TS-E and MFFG-TS-V outperformed other MFFG-based models, again indicating the effectiveness of the triple-stage training method.

Then, we compare the proposed methods with popular pre-trained models. Given that there are no directly available pre-trained language model baselines for video-based multimodal cross-lingual summarization, we first employ multilingual pre-trained models mT5[6] and mBART for text-only baselines. Then, we follow [5], [26] to construct VG-mT5 and VG-mBART, which adopt the forget gate fusion strategy [3] to integrate video information into mT5 and mBART for the MCLS task. VG-mT5/mBART and mT5/mBART are fine-tuned with the How2-MCLS data. In addition, we also observed the recent LLMs performance, using GPT-3.5 [53] for text-only baselines and GPT-4o[7] and Video-LLaMA [71] for multimodal baselines, where the prompts are shown in Table VII. Note that since most of the original videos can not be obtained due to invalid download links, we only evaluated the available samples on video LLMs. As can be observed from the experiment results

[6][Online], Available: https://huggingface.co/google/mt5-base

[7][Online], Available: https://cookbook.openai.com/examples/gpt4o/introduction_to_gpt4o

TABLE VII
THE PROMPT TEMPLATE OF THE LARGE LANGUAGE MODELS

LLM	Approach	Prompt
GPT-3.5	vanilla prompt (P)	Please summarize the following text in less than three sentences in English: [text]
	in-context learning (ICL)	The input article is: [text 1], and the summary is: [summary 1]. The input article is: [text 2], and the summary is: [summary 2]. The input article is: [text], and the summary is?
GPT-4o / Video-LLaMA	vanilla prompt	The video transcript text is: [text]. Please summarize this video in less than three sentences in English.

TABLE VIII
EXPERIMENT RESULTS ON PRETRAINED LANGUAGE MODELS

Method	V+Pt2En			V+Zh2En		
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-1	ROUGE-2	ROUGE-L
mT5	46.11	22.11	39.31	44.76	20.88	38.36
mBART	50.34	28.56	43.97	48.31	25.89	42.29
VG-mT5	47.35	23.80	41.24	45.67	21.76	39.09
VG-mBART	50.20	28.30	44.38	47.46	25.08	41.45
GPT-3.5 (P)	25.44	4.46	15.65	24.16	4.15	15.34
GPT-3.5 (ICL)	28.26	7.09	19.26	29.02	6.78	18.63
GPT-4o	28.53	4.26	18.22	28.76	4.58	17.70
Video-LLaMA	20.96	1.21	14.44	19.15	1.54	12.64
Pipe-Video-LLaMA	22.19	1.81	14.74	21.67	2.07	14.14
VDF	46.06	24.37	40.50	46.49	24.80	40.84
VDF-TS-V*	54.02	35.49	48.90	54.07	35.39	49.03

in Table VIII: 1) The proposed triple-stage training method, VDF-TS-V*, still exhibited an obvious advantage due to the transfer of prior MS knowledge and brings an absolute improvement of more than 8 ROUGE-L points on the VDF. 2) VG-mBART/mT5 and mBART/mT5 displayed decent performance due to their larger model architecture and extensive multilingual pre-training, surpassing the multimodal backbone VDF trained directly with How2-MCLS data. In particular, compared to mBART, mT5 offers more performance improvements in integrating video information. For example, VG-mT5 achieves a 1.93 ROUGE-L score increase than mT5. 3) GPT and VideoLLaMA performed moderately on both vanilla prompt and in-context learning strategies compared to supervised models. The observation reveals that despite LLM's robust unsupervised language understanding and generation capabilities, targeted training on domain-specific data remains essential.

We also constructed a unified MCLS model to observe its performance in both Portuguese-English and Chinese-English scenarios, where a multilingual encoder is adopted to performed knowledge distillation for both language pairs simultaneously. The experimental results are shown in Table IX. Compared to separate training, joint training achieves similar or improved performance. This improvement can be attributed to the fact that the training data for Chinese and Portuguese are similar in amount and most of them are parallel, allowing them to benefit from each other with less negative interference.

TABLE IX
PERFORMANCE OF USING AN UNIFIED MULTILINGUAL ENCODER FOR THE PROPOSED METHODS

Method	V+Pt2En & V+Zh2En					
VDF-TS-E	52.22	32.90	47.41	53.00	33.76	47.85
VDF-TS-V	51.83	33.22	47.13	52.81	34.00	47.96
VDF-TS-E*	52.79	33.44	48.11	53.29	34.29	48.53
VDF-TS-V*	53.77	35.84	49.26	54.49	36.21	49.90

E. Ablation Analysis

We construct ablation experiments in the Pt2En MCLS scenario to demonstrate the effectiveness of the components in the proposed methods. For the VDF model, we remove the dual fusion structures, including the fusion forget encoder (FF-Enc) and cascaded fusion decoder (CF-Dec). Besides, for the triple-stage training methods, we construct the following two ablations: a) remove the video input and the video fusion structures; b) remove the knowledge distillation phase from the triple-stage training paradigm, and the encoder/vocab are trained from scratch in the third training stage. The results are illustrated in Table X, and we could draw the following conclusions: 1) Eliminating either FF-Enc or CF-Dec components causes performance degradation of the VDF model, which reveals that the two fusion components are effective. In particular, compared to CF-Dec, removing FF-Enc leads to more performance degradation of VDF. This indicates that the forget gate fusion structure in the encoder plays more role in model performance gain. 2) The model performance dramatically decreases after removing the video input and video-related structures, which demonstrates the importance of multimodal information. 3) The proposed knowledge distillation mechanism brings 1.77 absolute ROUGE-L points promotion for VDF-TS-E, as the KD-induced triple-stage training method

TABLE X
ABLATION RESULTS OF THE PROPOSED METHODS

Method	ROUGE-1	ROUGE-2	ROUGE-L
VDF	46.06	24.37	40.50
w/o FF-Enc	44.53	22.69	39.61
w/o CF-Dec	45.63	23.12	40.12
VDF-TS-E	52.16	33.31	47.18
w/o KD	50.46	31.16	45.41
w/o video	47.40	27.23	42.14
VDF-TS-V	51.75	33.14	46.95
w/o KD	51.15	32.11	46.22
w/o video	47.14	27.35	41.96

TABLE XI
EFFECT OF MULTIMODAL FUSION AT DIFFERENT LAYERS

Enc \ Dec	Early fusion			Late fusion		
Early fusion	47.28	25.77	41.30	46.52	24.92	40.92
Late fusion	46.06	24.37	40.50	45.17	23.30	39.37

could effectively transfer the prior knowledge from the well-informed teacher model to the student, thus achieving better performances.

Considering that multimodal fusion occurs in both the encoder and decoder of the model, we set up the early fusion (at the first layer) and late fusion (at the last layer) in the encoder and decoder respectively to observe the impact of fusion at different parts of the model. As can be seen from experimental results in Table XI, whether in the encoder or decoder, the early fusion tends to achieve better performance compared to the late fusion. Although in our original setup, the encoder uses the late fusion and the decoder uses the early fusion, this was mainly to preserve more pure text encoder layers for cross-language knowledge distillation.

F. Knowledge Distillation Analysis

We explore the impact of the triple-stage training method by varying the knowledge distillation in different encoder layers. When the experiment is conducted on the 0-th encoder layer, it is equivalent to distilling the knowledge at the vocab-level. We conduct the experiment in the Video+Pt2En situation. From the experiment results in Table XII, we could observe that: (1) No matter the knowledge distillation is performed in which encoder sub-layer, the proposed methods outperform those without distillation, which illustrates the effectiveness of the knowledge distillation module. (2) Better experiment results are obtained in the middle encoder layer (layer = 2). (3) On the whole, the performance of each layer based on LAWD is better than that of adaptive pooling distillation. This indicates the advantages of LAWD in preserving the original language feature structure and reducing information loss compared to the adaptive pooling distillation.

TABLE XII
KNOWLEDGE DISTILLATION ON DIFFERENT ENCODER SUB-LAYERS

Method	ROUGE-1	ROUGE-2	ROUGE-L
VDF-TS-E (E*) w/o KD	50.46	31.16	45.41
VDF-TS-V (V*) w/o KD	51.15	32.11	46.22
Enc layer=4, VDF-TS-E	52.16	33.31	47.18
Enc layer=3, VDF-TS-E	52.48	33.62	47.36
Enc layer=2, VDF-TS-E	52.69	34.00	47.90
Enc layer=1, VDF-TS-E	52.64	33.86	47.86
Enc layer=0, VDF-TS-V	51.75	33.14	46.95
Enc layer=4, VDF-TS-E*	52.63	34.15	48.03
Enc layer=3, VDF-TS-E*	53.41	34.20	48.49
Enc layer=2, VDF-TS-E*	53.46	34.27	48.40
Enc layer=1, VDF-TS-E*	53.61	34.03	48.32
Enc layer=0, VDF-TS-V*	54.02	35.49	48.90

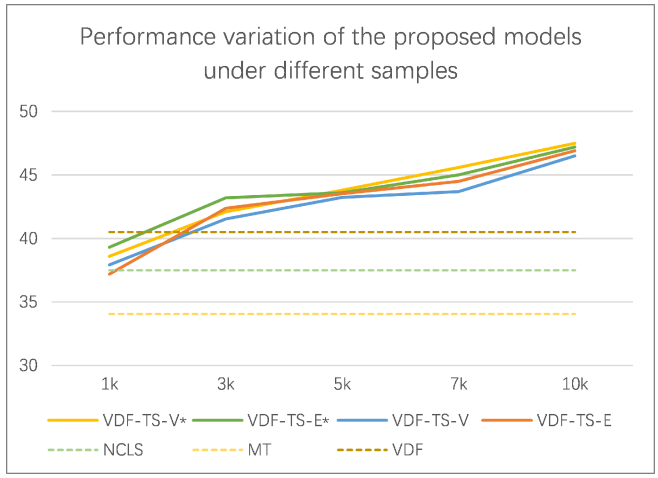

Fig. 7. ROUGE-L performance variations of the proposed models under different MCLS training samples. The dotted lines in the figure illustrate the models trained on the whole How2-MCLS dataset, and the solid lines are the results of the proposed triple-stage model with different sample sizes.

TABLE XIII
ZERO-SHOT PERFORMANCE OF THE PROPOSED METHODS

Method	ROUGE-1	ROUGE-2	ROUGE-L
VDF-TS-V (0-shot)	36.81	16.31	32.13
VDF-TS-E (0-shot)	40.48	20.46	35.90
VDF-TS-V* (0-shot)	42.10 (↑5.29)	21.57 (↑5.26)	37.13 (↑5.00)
VDF-TS-E* (0-shot)	47.82 (↑7.34)	26.62 (↑6.16)	42.65 (↑6.75)

G. Triple-Stage Training in Low-Resource Scenario

To investigate the model performance in lower resource MCLS scenarios, we conduct experiments under the Video+Pt2En circumstance by reducing the MCLS samples. To be specific, in the first phase of the proposed triple-stage training method, we still leverage the complete MS data, but the number of samples from MCLS in the second and third stages is reduced. The experimental results are shown in Fig. 7. We could observe that using as few as 1 k samples in the How2-MCLS dataset, VDF-TS-V (V*) and VDF-TS-E (E*) achieve comparable or better performance than the unimodal models (e.g., NCLS and MT), which utilize the 13 k samples for training. Meanwhile,

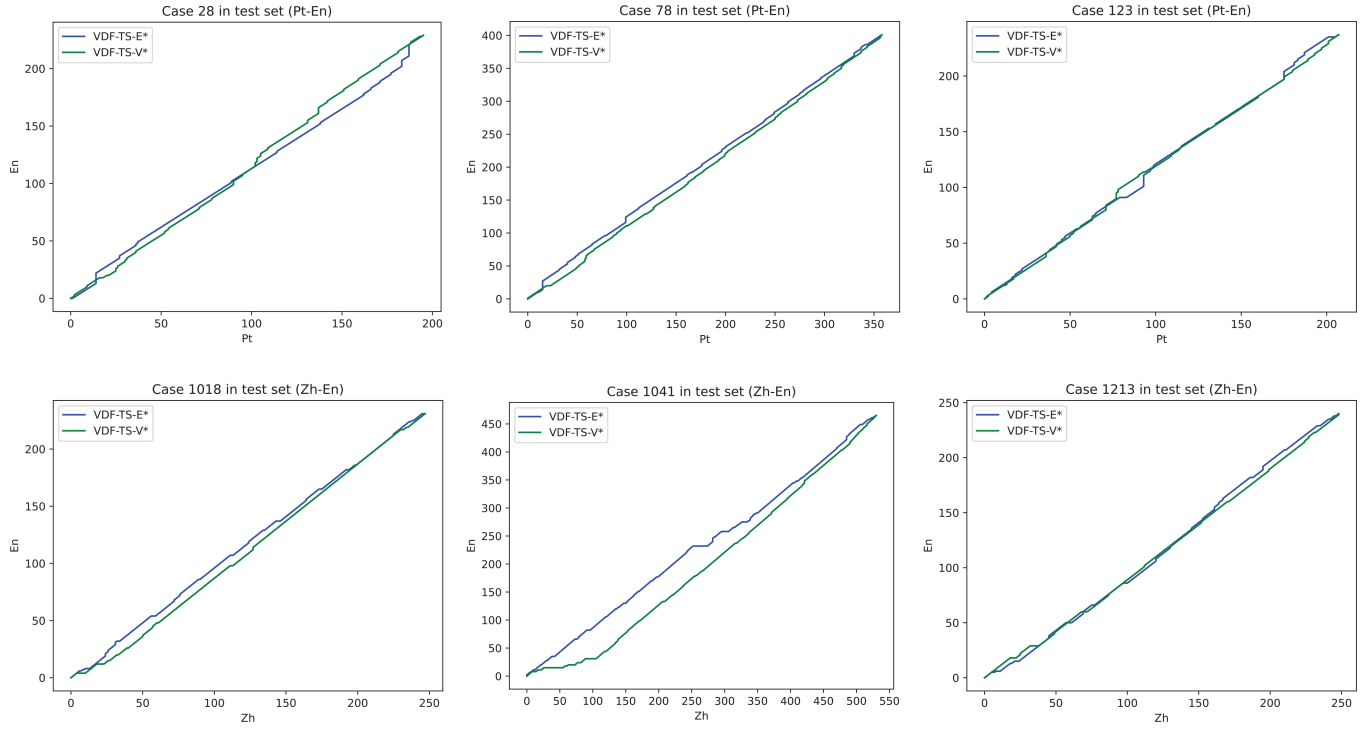

Fig. 8. Visualization of language-adaptive warping distillation (LAWD). These cases are taken from the How2-MCLS test set. Case 28, 78, and 123 in Pt-En MCLS scenario are displayed. Case 1018, 1041, and 1213 in Zh-En MCLS scenario are displayed.

with only leveraging 3 k How2-MCLS data, VDF-TS-V (V*) and VDF-TS-E (E*) outperform baseline model VDF that is trained with the full amount of How2-MCLS data. Such findings illustrate the superiority of the proposed triple-stage training method in the low-resource MCLS situation.

Additionally, we conducted experiments without using the third-stage MCLS data for training, only keeping the first and second stages of the triple-stage training framework to observe its performance in the zero-shot situation, which aims to better observe the ability of pure cross-lingual knowledge distillation components to transfer model knowledge. As can be seen from the experimental results in Table XIII, LAWD demonstrated great performance transfer capabilities. Compared to the performance of different distillation methods under complete triple-stage training, LAWD's advantage in the zero-shot scenario was more pronounced, showing an improvement of 5.00 ROUGE-L points on vocab-level distillation and 6.75 ROUGE-L points on encoder-level distillation than the adaptive pooling distillation. This demonstrates the superiority of LAWD in directly performing distillation while maintaining the shape and position of unequal-length cross-lingual parallel feature sequences unchanged. We also found that encoder-level distillation outperforms vocab-level distillation. We speculate that without MCLS data training, vocab-level distillation can only alter the knowledge of vocab embeddings, and the encoder still stores a large amount of English knowledge, leading to insufficient learned representations. In contrast, encoder-level distillation leverages more parameters to comprehensively learn cross-lingual alignment and replaces more knowledge of the high-resource language encoder with the low-resource language encoder, thus achieving better performance.

H. Visualization Analysis

In this section, we visualize LAWD to intuitively observe how source and target language sequences of different lengths output by the teacher and student models establish alignment without changing the feature shape during the distillation process. Specifically, in the teacher-student cost-matrix described in Section III-D2, we take the two transcript texts of parallel bilinguals in the How2-MCLS test set as input, and extract the teacher-student cost matrix in the second-stage distillation model in the triple-stage training method. On the teacher-student cost matrix, we find the minimum cost path through the dynamic programming ((10)), which represents the potential alignment relationship learned by the source and target language sequence. We assume that paths closer to the diagonal perform better. We take three cases in the Pt-En and Zh-En scenarios respectively, and draw the minimum cost path curves of VDF-TS-V* and VDF-TS-E* on the coordinate axes for each case, where the horizontal and vertical coordinates represent the length of the feature sequence output by the teacher and student models. The visualization results are shown in Fig. 8. It can be seen that the minimum cost path is generally close to the diagonal. In the figure, the oblique lines on the path indicate that after the current tokens of the two sequences are aligned, they start to align the next token. The horizontal and vertical lines indicate that the alignment of the source and target language sequences is distorted, that is, the token in one sequence is waiting for a

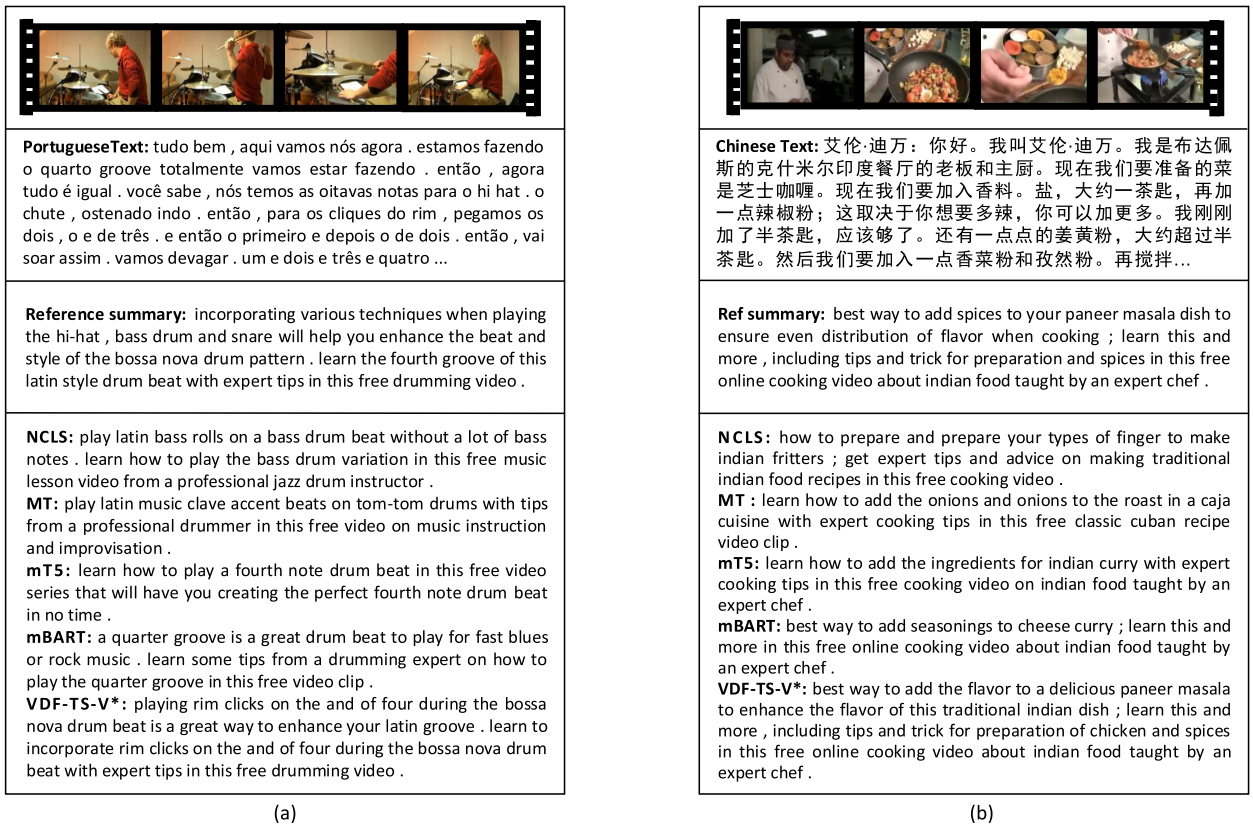

Fig. 9. Case study. Case 28 (Pt2En) and Case 159 (Zh2En) from the How2-MCLS test set are displayed.

TABLE XIV
FAILURE CASE ANALYSIS

Case	Generation	Ground truth
a	in skateboarding , a front side 50-50 should n't be performed to ollie on a flat surface . avoid common mistakes when doing a front side 50-50 with tips from a sponsored skateboarder in this free video on skateboarding tricks .	get skateboarding trick tips for avoiding common mistakes with the fakie smith stall in this free skateboarding video .
b	the half lotus pose in yoga is a great way to open the hips and relieve stress . learn assisted yoga poses like the standing forward fold from a certified instructor in this free yoga video .	keep hips on the wall when doing a forward fold pose in standing yoga . learn how to do a forward fold pose in this free standing yoga video from a fitness and yoga instructor .

more distant token in another sequence. This distortion enables the source and target language sequences to fully learn potential alignment relationships at different lengths.

I. Case Study

Several examples are provided for an intuitive observation of model performance. As shown in Fig. 9, we display some key screenshots from the video, the transcript text, reference summaries, and cross-lingual summaries generated by different models. From the generated results, we can observe that: Predictions based on pure video modality or pure text modality are not very accurate. For instance, in example (b), the country is not mentioned in the video, but it is mentioned in the text (Indian cuisine). The model MT, which only uses the video modality, predicted the Indian cuisine as Cuban cuisine, highlighting the importance of multimodal fusion. The proposed methods leverage cross-lingual distillation to transfer knowledge from MS models trained with abundant resources, and achieves accurate and comprehensive content. For example, in case (a), VDF-TS-V* not only predicts details like "bossa nova" and "latin groove" that were not mentioned in the transcript, but also generates some details from the original transcript in a cross-lingual format that were not mentioned in the summary, such as "rim clicks" (corresponding to "cliques do rim" in Portuguese). This observation verifies the role of the proposed method in establishing cross-lingual alignment.

J. Failure Case Analysis

This section lists some failure cases generated by the proposed triple-stage training method. Although the overall summaries

are fluent, a few generated results exhibit hallucinations [72], i.e., generating unsupported or fabricated content, as shown in Table XIV. For example, in case a, the ground truth refers to the skateboarding action "fakie smith stall", and VDF-TS-V* predicts a similar action "front side 50-50"; In case b, although VDF-TS-V* predicts the yoga posture "forward fold" mentioned by ground truth but also hallucinates the "half lotus pose." These hallucination contents appear with a certain frequency in the training data of the monolingual MS model. We speculate that the proposed triple-stage training model, while leveraging the prior knowledge from the monolingual MS model to enhance performance, excessively relies on these pretrained parameters, paying insufficient attention to new information in the MCLS context data, thereby generating hallucinations [73]. Therefore, leveraging MS prior knowledge to enhance the performance of the MCLS model while suppressing potential hallucinations generated by prior knowledge is a direction that we still need to improve in the future.

V. Conclusion

In this work, oriented the need of non-native viewers to understand cross-lingual videos, we introduce the study of multimodal cross-lingual summarization for videos (MCLS), which assists non-native viewers with native language summaries generated from non-native videos. Concretely, we propose a knowledge distillation induced triple-stage training method, which aims to transfer knowledge from resource-rich MS models to enhance MCLS model learning. In triple-stage training, a video-guided dual fusion network (VDF) is designed to be used as a backbone to integrate multimodal and cross-lingual information to generate summaries; moreover, two knowledge distillation strategies, adaptive pooling distillation and language-adaptive warping distillation (LAWD) with different distillation objects (encoder-level and vocab-level), are proposed for constructing cross-language alignment between MS and MCLS models. In particular, LAWD can directly conduct knowledge distillation on parallel cross-language feature sequences of unequal length without changing the language feature structure. Extensive experimental results illustrate the effectiveness of the proposed method.

In the future, leveraging MS prior knowledge to enhance MCLS model performance while suppressing potential hallucinations still needs to be improved. Additionally, due to the high cost of manually annotated MCLS data, using noisy summary data from automated machine translation to refine model performance is also a worthy research topic.

References

[1] J. Libovickỳ, S. Palaskar, S. Gella, and F. Metze, "Multimodal abstractive summarization for open-domain videos," in *Proc. Workshop Visually Grounded Interaction Lang.*, 2018, pp. 1–8.

[2] S. Palaskar, J. Libovický, S. Gella, and F. Metze, "Multimodal abstractive summarization for how2 videos," in *Proc. 57th Conf. Assoc. Comput. Linguistics*, 2019, pp. 6587–6596.

[3] N. Liu, X. Sun, H. Yu, W. Zhang, and G. Xu, "Multistage fusion with forget gate for multimodal summarization in open-domain videos," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 1834–1845.

[4] N. Liu, X. Sun, H. Yu, F. Yao, G. Xu, and K. Fu, "Abstractive summarization for video: A revisit in multistage fusion network with forget gate," *IEEE Trans. Multimedia*, vol. 25, pp. 3296–3310, 2023.

[5] T. Yu, W. Dai, Z. Liu, and P. Fung, "Vision guided generative pre-trained language models for multimodal abstractive summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 3995–4007.

[6] M. Li, X. Chen, S. Gao, Z. Chan, D. Zhao, and R. Yan, "VMSMO: Learning to generate multimodal summary for video-based news articles," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 9360–9369.

[7] P. Tang, K. Hu, L. Zhang, J. Luo, and Z. Wang, "TLDW: Extreme multimodal summarisation of news videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 3, pp. 1469–1480, Mar. 2024.

[8] T. B. Amin and I. Mahmood, "Speech recognition using dynamic time warping," in *Proc. IEEE 2nd Int. Conf. Adv. Space Technol.*, 2008, pp. 74–79.

[9] M. Cuturi and M. Blondel, "Soft-DTW: A differentiable loss function for time-series," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2017, pp. 894–903.

[10] R. Sanabria et al., "How2: A large-scale dataset for multimodal language understanding," 2018, *arXiv:1811.00347*.

[11] N. Liu et al., "Assist non-native viewers: Multimodal cross-lingual summarization for how2 videos," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 6959–6969.

[12] X. Qian, M. Li, Y. Ren, and S. Jiang, "Social media based event summarization by user–text–image co-clustering," *Knowl.-Based Syst.*, vol. 164, pp. 107–121, 2019.

[13] M. Li, L. Zhang, H. Ji, and R. J. Radke, "Keep meeting summaries on topic: Abstractive multi-modal meeting summarization," in *Proc. 57th Conf. Assoc. Comput. Linguistics*, 2019, pp. 2190–2196.

[14] A. Khullar and U. Arora, "MAST: Multimodal abstractive summarization with trimodal hierarchical attention," in *Proc. 1st Int. Workshop Natural Lang. Process. Beyond Text*, 2020, pp. 60–69.

[15] H. Li, J. Zhu, C. Ma, J. Zhang, and C. Zong, "Multi-modal summarization for asynchronous collection of text, image, audio and video," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1092–1102.

[16] A. Jangra, A. Jatowt, M. Hasanuzzaman, and S. Saha, "Text-image-video summary generation using joint integer linear programming," in *Proc. 42nd Eur. Conf. Adv. Inf. Retrieval*, Springer, 2020, pp. 190–198.

[17] J. Zhu, H. Li, T. Liu, Y. Zhou, J. Zhang, and C. Zong, "MSMO: Multimodal summarization with multimodal output," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 4154–4164.

[18] J. Chen and H. Zhuge, "Abstractive text-image summarization using multimodal attentional hierarchical RNN," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 4046–4056.

[19] H. Li, J. Zhu, T. Liu, J. Zhang, and C. Zong, "Multi-modal sentence summarization with modality attention and image filtering," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 4152–4158.

[20] H. Li, P. Yuan, S. Xu, Y. Wu, X. He, and B. Zhou, "Aspect-aware multimodal summarization for chinese e-commerce products," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 8188–8195.

[21] Q. Lu, X. Ye, and C. Zhu, "MTCA: A multimodal summarization model based on two-stream cross attention," in *Proc. IEEE 2nd Int. Conf. Comput. Sci. Electron. Inf. Eng. Intell. Control Technol.*, 2022, pp. 594–601.

[22] J. Im, M. Kim, H. Lee, H. Cho, and S. Chung, "Self-supervised multimodal opinion summarization," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics, 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 388–403.

[23] J. Zhu, Y. Zhou, J. Zhang, H. Li, C. Zong, and C. Li, "Multimodal summarization with guidance of multimodal reference," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 9749–9756.

[24] D. Lin, L. Jing, X. Song, M. Liu, T. Sun, and L. Nie, "Adapting generative pretrained language model for open-domain multimodal sentence summarization," in *Proc. 46th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2023, pp. 195–204.

[25] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7871–7880.

[26] Y. Liang, F. Meng, J. Xu, J. Wang, Y. Chen, and J. Zhou, "Summary-oriented vision modeling for multimodal abstractive summarization," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics*, 2023, pp. 2934–2951.

[27] Z. Xie et al., "SimMIM: A simple framework for masked image modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9643–9653.

[28] S. Xiaorui, "MCLS: A large-scale multimodal cross-lingual summarization dataset," in *Proc. 22nd Chin. Nat. Conf. Comput. Linguistics*, 2023, pp. 862–874.

[29] Y. Liang, F. Meng, J. Wang, J. Xu, Y. Chen, and J. Zhou, "D2TV: Dual knowledge distillation and target-oriented vision modeling for many-to-many multimodal summarization," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2023, pp. 14 910–14 922.
[30] Z. Li, Y. Guo, K. Wang, Y. Wei, L. Nie, and M. Kankanhalli, "Joint answering and explanation for visual commonsense reasoning," *IEEE Trans. Image Process.*, vol. 32, pp. 3836–3846, 2023.
[31] X. Shang, Z. Yuan, A. Wang, and C. Wang, "Multimodal video summarization via time-aware transformers," in *Proc. ACM Multimedia Conf.*, 2021, pp. 1756–1765.
[32] Y. K. Atri, S. Pramanick, V. Goyal, and T. Chakraborty, "See, hear, read: Leveraging multimodality with guided attention for abstractive text summarization," *Knowl.-Based Syst.*, vol. 227, pp. 107–152, 2021.
[33] C. Zhang, Z. Zhang, J. Li, Q. Liu, and H. Zhu, "CtnR: Compress-then-reconstruct approach for multimodal abstractive summarization," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, 2021, pp. 1–8.
[34] J. Qiu et al., "MHMS: Multimodal hierarchical multimedia summarization," 2022, *arXiv:2204.03734*.
[35] M. Krubinski and P. Pecina, "MLASK: Multimodal summarization of video-based news articles," in *Proc. Assoc. Comput. Linguistics*, 2023, pp. 880–894.
[36] J. Wang et al., "A survey on cross-lingual summarization," *Trans. Assoc. Comput. Linguistics*, vol. 10, pp. 1304–1323, 2022.
[37] A. Leuski, C. Lin, L. Zhou, U. Germann, F. J. Och, and E. H. Hovy, "Cross-lingual C*ST*RD: English access to hindi information," *ACM Trans. Asian Lang. Inf. Process.*, vol. 2, no. 3, pp. 245–269, 2003.
[38] X. Wan, "Using bilingual information for cross-language document summarization," in *Proc. Conf. 49th Annu. Meeting Assoc. Comput. Linguistics Hum. Lang. Technol.*, 2011, pp. 1546–1555.
[39] C. Orasan and O. A. Chiorean, "Evaluation of a cross-lingual romanian-english multi-document summariser," in *Proc. Int. Conf. Lang. Resour. Eval.*, 2008, pp. 2114–2119.
[40] J. Zhang, Y. Zhou, and C. Zong, "Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 10, pp. 1842–1853, Oct. 2016.
[41] J. Yao, X. Wan, and J. Xiao, "Phrase-based compressive cross-language summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 118–127.
[42] B. Lv et al., "DSP: Discriminative soft prompts for zero-shot entity and relation extraction," in *Proc. Assoc. Comput. Linguistics*, 2023, pp. 5491–5505.
[43] X. Wan, H. Li, and J. Xiao, "Cross-language document summarization based on machine translation quality prediction," in *Proc. 48th Annu. Meeting Assoc. Comput. Linguistics*, 2010, pp. 917–926.
[44] F. Ladhak, E. Durmus, C. Cardie, and K. Mckeown, "WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2020, pp. 4034–4048.
[45] J. Zhu et al., "NCLS: Neural cross-lingual summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 3052–3062.
[46] J. Zhu, Y. Zhou, J. Zhang, and C. Zong, "Attend, translate and summarize: An efficient method for neural cross-lingual summarization," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 1309–1321.
[47] T. T. Nguyen and A. T. Luu, "Improving neural cross-lingual abstractive summarization via employing optimal transport distance for knowledge distillation," in *Proc. 36th AAAI Conf. Artif. Intell.*, 2022, pp. 11 103–11 111.
[48] P. Li, Z. Zhang, J. Wang, L. Li, A. Jatowt, and Z. Yang, "ACROSS: An alignment-based framework for low-resource many-to-one cross-lingual summarization," in *Proc. Assoc. Comput. Linguistics*, 2023, pp. 2458–2472.
[49] S. Takase and N. Okazaki, "Multi-task learning for cross-lingual abstractive summarization," in *Proc. 13th Lang. Resour. Eval. Conf.*, 2022, pp. 3008–3016.
[50] Y. Liang et al., "A variational hierarchical model for neural cross-lingual summarization," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, 2022, pp. 2088–2099.
[51] R. Xu, C. Zhu, Y. Shi, M. Zeng, and X. Huang, "Mixed-lingual pre-training for cross-lingual summarization," in *Proc. 1st Conf. Asia-Pacific Chapter Assoc. Comput. Linguistics, 10th Int. Joint Conf. Natural Lang. Process.*, 2020, pp. 536–541.
[52] J. Wang et al., "Zero-shot cross-lingual summarization via large language models," in *Proc. 4th New Front. Summarization Workshop*, 2023, pp. 12–23.
[53] L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 27 730–27 744.
[54] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet?," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6546–6555.
[55] W. Kay et al., "The kinetics human action video dataset," 2017, *arXiv: 1705.06950*.
[56] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
[57] M. Li, P.-Y. Huang, X. Chang, J. Hu, Y. Yang, and A. Hauptmann, "Video pivoting unsupervised multi-modal machine translation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3918–3932, Mar. 2023.
[58] Z. Zhang et al., "Universal multimodal representation for language understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 9169–9185, Jul. 2023.
[59] B. Lv, X. Liu, K. Wei, P. Luo, and Y. Yu, "TAeKD: Teacher assistant enhanced knowledge distillation for closed-source multilingual neural machine translation," in *Proc. Joint Int. Conf. Comput. Linguistics Lang. Resour. Eval.*, 2024, pp. 15 530–15 541.
[60] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8717–8727, Dec. 2022.
[61] K. Saenko, K. Livescu, J. Glass, and T. Darrell, "Multistream articulatory feature-based models for visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 9, pp. 1700–1707, Sep. 2009.
[62] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Workshop Text Summarization Branches Out*, 2004, pp. 74–81.
[63] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2002, pp. 311–318.
[64] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. Summarization*, 2005, pp. 65–72.
[65] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4566–4575.
[66] J. Wang, W. Xu, Q. Wang, and A. B. Chan, "On distinctive image captioning via comparing and reweighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2088–2103, Feb. 2023.
[67] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1412–1421.
[68] J. Zhu et al., "NCLS: Neural cross-lingual summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 3054–3064.
[69] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8739–8748.
[70] Y. Liu et al., "Multilingual denoising pre-training for neural machine translation," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 726–742, 2020.
[71] H. Zhang, X. Li, and L. Bing, "Video-LLaMA: An instruction-tuned audio-visual language model for video understanding," in *Proc. Conf. Empirical Methods Natural Lang. Process. Syst. Demonstrations*, 2023, pp. 543–553.
[72] L. Huang et al., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," 2023, *arXiv:2311.05232*.
[73] W. Shi, X. Han, M. Lewis, Y. Tsvetkov, L. Zettlemoyer, and S. W.-T. Yih, "Trusting your evidence: Hallucinate less with context-aware decoding," 2023, *arXiv:2305.14739*.

Nayu Liu received the BSc degree from Xidian University, Xian, China, in 2018, and the PhD degree from the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China, in 2023. He is currently an assistant professor with the School of Computer Science and Technology, Tiangong University, Tianjin, China. His research interests include deep learning, natural language processing, and multi-modal learning.

Kaiwen Wei (Member, IEEE) received the BSc degree from Chongqing University, Chongqing, China, in 2019. He is currently working toward the PhD degree with Aerospace Information Innovation Institute, Chinese Academy of Sciences, China. His research interests include deep learning, natural language processing and information extraction.

Yong Yang (Senior Member, IEEE) received the PhD degree from Xi'an Jiaotong University, Xi'an, China, in 2005. From 2009 to 2010, he was a postdoctoral research fellow with Chonbuk National University, Jeonju, South Korea. He is currently a distinguished professor with the School of Computer Science and Technology, Tiangong University, Tianjin, China. His current research interests include image processing, pattern recognition, and deep learning. He is an editor of the *KSII Transactions on Internet and Information Systems*.

Jianhua Tao (Senior Member, IEEE) received the MS degree from Nanjing University, Nanjing, China, in 1996, and the PhD degree from Tsinghua University, Beijing, China, in 2001. He is currently a professor with Department of Automation, Tsinghua University, Beijing, China. He has authored or coauthored more than 300 papers on major journals and proceedings including the *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *IEEE Transactions on Affective Computing*, *IEEE Transactions on Image Processing*, *IEEE Transactions on Systems, Man, and Cybernetics*, *Information Fusion*, etc. His current research interests include speech recognition and synthesis, affective computing, and pattern recognition. He is the board member of ISCA, the chairperson of ISCA SIG-CSLP, the chair or program committee member for several major conferences, including Interspeech, ICPR, ACII, ICMI, ISCSLP, etc. He was the subject editor for the Speech Communication, and is an associate editor for *Journal on Multimodal User Interface* and *International Journal on Synthetic Emotions*. He was the recipient of several awards from important conferences, including Interspeech, NCMMSC, etc.

Xian Sun (Senior Member, IEEE) received the BSc degree from Beihang University, Beijing, China, in 2004, and the MSc and PhD degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, in 2006 and 2009, respectively. He is currently a professor with Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include computer vision, geospatial data mining, and remote-sensing image understanding.

Fanglong Yao (Member, IEEE) received the BSc degree from Inner Mongolia University, Hohhot, China, in 2017, and the PhD degree from the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China, in 2022. He is currently a postdoctoral researcher with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include cognitive intelligence, embodied intelligence, and swarm intelligence, concentrating on multiagent learning, multimodal fusion and reasoning, 3-D scene understanding, and spatiotemporal data analysis.

Hongfeng Yu received the BSc and MSc degree from Peking University, Beijing, China, in 2013 and 2016 respectively, and the PhD degree from the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China, in 2023. He is currently a research assistant with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing. His research interests include deep learning and natural language processing.

Li Jin received the BS degree from Xidian University, Xi'an, China, in 2012 and the PhD degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2017. He is currently an associate professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include machine learning, knowledge graph and geographic information processing.

Zhao Lv (Member, IEEE) received the PhD degree in computer application technology from Anhui University, Hefei, China, in 2011. He was a visiting scholar with the University of Utah, Salt Lake City, USA, from 2017 to 2018. He is currently a professor with the School of Computer Science and Technology, Anhui University, Hefei, China. His research interests include intelligent information processing and pattern recognition regarding biomedical signals (EEG, EOG, etc.) as well as speech signal processing.

Cunhang Fan (Member, IEEE) received the BS degree from the Beijing University of Chemical Technology (BUCT), Beijing, China, in 2016 and the PhD degree with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2021. He is currently an associate professor with the School of Computer Science and Technology, Anhui University, Heifei, China. His current research interests include speech enhancement, fake speech detection, speech recognition and speech processing.