Assist Non-native Viewers: Multimodal Cross-Lingual Summarization for How2 Videos

Nayu Liu[1,2]*, **Kaiwen Wei**[1,2]*, **Xian Sun**[1,2], **Hongfeng Yu**[1], **Fanglong Yao**[1]†, **Li Jin**[1], **Zhi Guo**[1], **Guangluan Xu**[1]
[1]Key Laboratory of Network Information System Technology, Aerospace Information Research Institute, Chinese Academy of Sciences
[2]School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences
{liunayu18,weikaiwen19,yaofanglong17}@mails.ucas.ac.cn

Abstract

Multimodal summarization for videos aims to generate summaries from multi-source information (videos, audio transcripts), which has achieved promising progress. However, existing works are restricted to monolingual video scenarios, ignoring the demands of non-native video viewers to understand the cross-language videos in practical applications. It stimulates us to propose a new task, named **M**ultimodal **C**ross-**L**ingual **S**ummarization for videos (MCLS), which aims to generate cross-lingual summaries from multimodal inputs of videos. First, to make it applicable to MCLS scenarios, we conduct a Video-guided Dual Fusion network (VDF) that integrates multimodal and cross-lingual information via diverse fusion strategies at both encoder and decoder. Moreover, to alleviate the problem of high annotation costs and limited resources in MCLS, we propose a triple-stage training framework to assist MCLS by transferring the knowledge from monolingual multimodal summarization data, which includes: 1) multimodal summarization on sufficient prevalent language videos with a VDF model; 2) knowledge distillation (KD) guided adjustment on bilingual transcripts; 3) multimodal summarization for cross-lingual videos with a KD induced VDF model. Experiment results on the reorganized How2 dataset show that the VDF model alone outperforms previous methods for multimodal summarization, and the performance further improves by a large margin via the proposed triple-stage training framework.

1 Introduction

Multimodal summarization (MS) for videos aims at integrating multimodal information such as videos and audio transcriptions 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, which benefits users from better understanding and accessing to verbose and obscure videos.

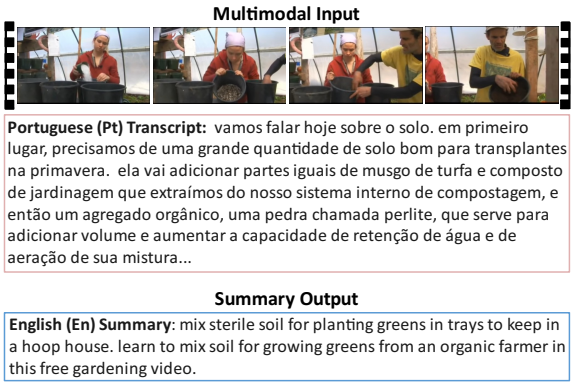

Figure 1: An example of the MCLS task, which seeks to generate a target language (e.g., English) summary based on the video and its transcript in the source language (e.g., Portuguese).

Palaskar et al. (2019) introduced the multimodal summarization for open-domain videos. Following the predefined task, former state-of-the-art multimodal summarization methods have achieved great outcomes. For example, Liu et al. (2020) proposed a multistage fusion network, and Shang et al. (2021) utilized time-aware multimodal transformers. However, 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 propose a new task: **M**ultimodal **C**ross-**L**ingual **S**ummarization 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 audio transcriptions in the source language.

*Equal contribution.
†Corresponding author.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 6959 - 6969

There are two challenging issues for MCLS: (1) *Messy multimodal and cross-lingual information*: compared with recent summarization methods in multimodal (Khullar and Arora, 2020; Zhu et al., 2018; Yu et al., 2021) or textual cross-lingual (Cao et al., 2020; Zhu et al., 2020) scenario, it is more challenging to effectively integrate multimodal and different language information simultaneously, and then generate summaries. (2) *Insufficient multimodal cross-lingual summarization data*: due to the relatively small number of bilingualist, building a high-quality multimodal cross-lingual summarization dataset is costly. The experiment results of directly training on such limited multimodal data are typically under satisfactory. There are several textual low-resource cross-lingual summarization researches that focus on leveraging external toolkits (Jiang et al., 2022) or pre-trained language models (PLMs) (Xu et al., 2020). However, since MCLS is a newly proposed task, there are no available tools or PLMs that can be directly adopted.

In this paper, to overcome the first issue in MCLS, 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. Furthermore, to solve the second one, we propose a triple-stage training framework, which leverages the knowledge of the multimodal summarization model trained with prevalent mono language to assist the low-resource MCLS generation. In the first stage, a VDF model is trained on sufficient multimodal summarization data in the target language. In the second stage, encoder- and vocab-level knowledge 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 knowledge distillation 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.

To simulate the MCLS, we reorganize the How2 dataset (Sanabria et al., 2018), a large-scale multimodal understanding dataset of open-domain videos, for two scenarios according to the cross-lingual video-summary data volume: 1) general MCLS: contains Portuguese videos with English summaries; 2) MS-augmented MCLS: on the basis of 1), it is supplemented with sufficient English video-summary data and a small amount of bilingual English-Portuguese transcripts. Experiment results show that the VDF model alone outperforms state-of-the-art methods in the general MCLS situation, and the performance further improves by a large margin via the proposed triple-stage training framework. What's more, with only 3k samples, our triple-stage training method outperforms strong baselines in general MCLS, which is trained with more than 10k samples. The contribution of this work could be summarized as follows:

1) We introduce a new task, named 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 the general MCLS scenario.

2) We propose a triple-stage training framework to alleviate the problem of limited resources in MCLS, where a knowledge distillation method is designed to drive a VDF model that benefits from sufficient MS data in the prevalent language.

3) We reorganize the How2 dataset to simulate MCLS. Experiment results illustrate that our methods achieves state-of-the-art performance on both general and MS-augmented MCLS scenarios[1].

2 Method

2.1 Overview

Given a video and its corresponding transcript 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). Simulated by the How2 dataset, we consider the following two MCLS scenarios:

(1) **General MCLS**: contains the Portuguese video to English summary data.

(2) **MS-augmented MCLS**: oriented to alleviate the problem of limited resources in MCLS, based on the data from (1), it is accompanied by sufficient English video-summary data and some parallel English-Portuguese transcripts.

Formally, let the input video representations be $V = (v_1, ..., v_k)$, where v_k is the feature vector extracted by a pre-trained model. Besides, assume we have the input English transcript $X = (x_1, ..., x_n)$ and Portuguese transcript $Y = (y_1, ..., y_m)$, which consist of n and m words. The output English

[1]The code and reorganized data are available at https://github.com/korokes/MCLS

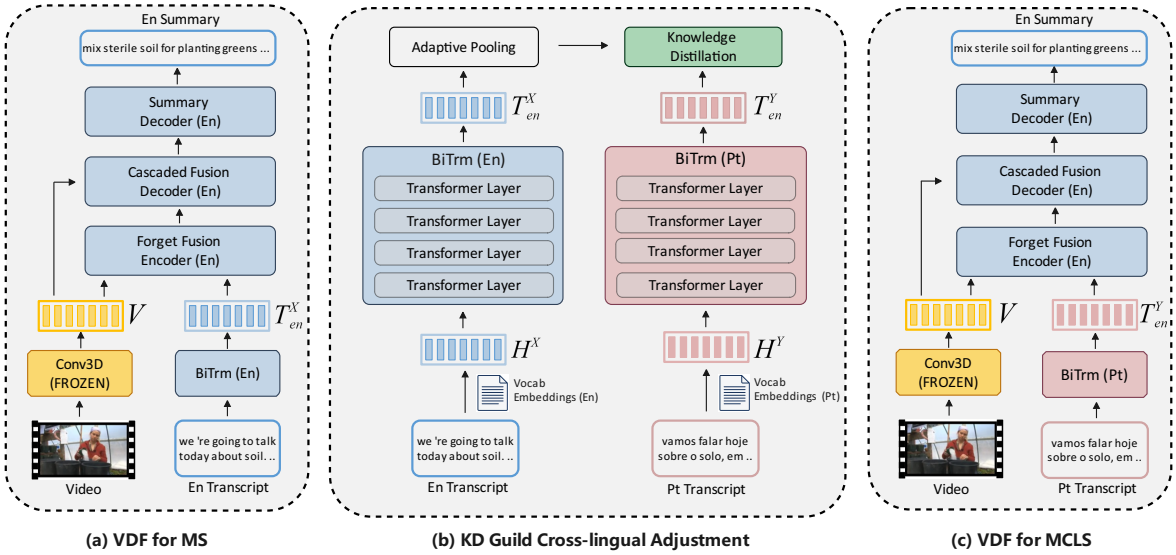

Figure 2: The proposed triple-stage training framework for MCLS, which includes: (1) training a VDF model for multimodal summarization (MS) on English video-summary corpus; (2) cross-lingual encoder adjustment via knowledge distillation (KD); (3) replacing the English encoder with the Portuguese one, 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.

summary could be denoted as a sequence of word tokens $S = (s_1, ..., s_l)$ consisting of several sentences. The task aims to predict the best summary sequence S by finding:

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

where θ is the set of trainable parameters. For the general MCLS scenario without the assistance of the English transcript, we set $X = \{None\}$.

In the following sections, we first show the details of the VDF model for general MCLS, and then describe our triple-stage training framework that utilizes a knowledge distillation method to drive a VDF model to alleviate the low-resource problem in the MS-augmented scenario.

2.2 Video-guided Dual Fusion Network

Fig. 2 (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. In the general MCLS scenario, we could directly train the VDF model on the given videos, Portuguese transcripts, and English summaries.

2.2.1 Video and Text Encoder

Encoding Video. The video encoding features $V = (v_1, ..., v_m)$ are extracted from every 16 nonoverlapping frames by a pretrained action recognition model: a ResNeXt-101 3D convolutional neural network (Hara et al., 2018) trained for recognizing 400 different human actions in the Kinetics dataset (Kay et al., 2017).

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

We add learnable position embeddings for video features.

Encoding Transcript. We apply a standard bidirectional transformer encoder (Vaswani et al., 2017) to get contextual text encoding features, where each layer is composed of a multi-head self-attention layer and a feedforward layer. Take Portuguese transcript Y as an example, the encoding process could be denoted by the following equation:

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

where "BiTrm" means the standard bidirectional transformer encoder.

2.2.2 Forget Fusion Encoder

Considering the source transcript text is lengthy and has much redundancy, forget fusion encoder first adopts the forget gate fusion (FGF) (Liu et al., 2020) module, which fuses video information to

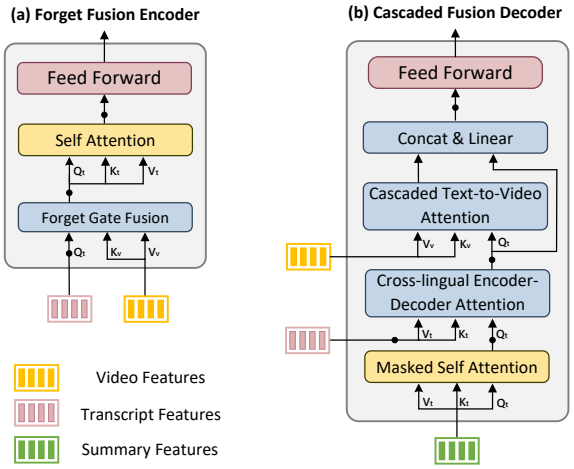

Figure 3: The multimodal feature fusion strategies at encoding and decoding in VDF, include (a) forget fusion encoder and (b) cascaded fusion decoder, respectively. "·" represents a layer normalization operation.

text while suppressing the flow of cross-modal noise. As illustrated in Fig. 3 (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}.

2.2.3 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. 3 (b), the decoder consists of three sets of cascaded scaled dot-product attention layers: first, the target English summaries are fed to a masked 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 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 vector 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:

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

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

2.3 Triple-stage Training Framework

To alleviate the problem of limited resources in the MS-augmented MCLS scenario, we further propose a triple-stage training framework, which transfers the knowledge of the model trained with sufficient English multimodal summarization data to the model under MCLS data. As illustrated in Fig. 2, it consists of the following three stages:

(1) Multimodal summarization for sufficient mono language videos: leveraging VDF as the backbone and training an English video to English summary model $\phi_{\mathcal{X}}$.

(2) Knowledge distillation (KD) from the prevalent language to the objective one through parallel bilingual transcripts, where two strategies are designed: a) encoder-level KD: transferring knowledge of the English encoder $\phi_{E_{\mathcal{X}}}$ in VDF to a new Portuguese encoder $\phi_{E_{\mathcal{Y}}}$; b) vocab-level KD: only transferring knowledge of the English vocab embedding table $\phi_{V_{\mathcal{X}}}$ in VDF to a new Portuguese vocab embedding table $\phi_{V_{\mathcal{Y}}}$.

(3) Multimodal summarization for cross-lingual videos: replacing the English encoder $\phi_{E_{\mathcal{X}}}$ (vocab $\phi_{V_{\mathcal{X}}}$) with the Portuguese encoder $\phi_{E_{\mathcal{Y}}}$ (vocab $\phi_{V_{\mathcal{Y}}}$) to form a new VDF model $\phi_{\mathcal{Y}}$. Then fine-tune $\phi_{\mathcal{Y}}$ on the Portuguese videos to generate English summaries.

2.4 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 framework to

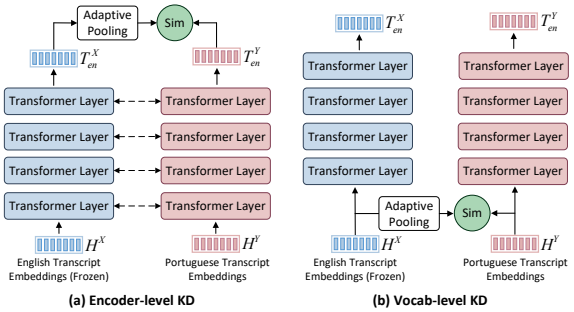

Figure 4: Illustration of the knowledge distillation (KD) guide cross-lingual adjustment, including: (a) encoder-level KD; (b) vocab-level KD.

transfer the knowledge from English (teacher) to Portuguese (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. After feeding to the encoder $\phi_{E_{\mathcal{X}}}$ of VDF $\phi_{\mathcal{X}}$, we could obtain the English encoder output T^X_{en}.

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. 4 (a), a new randomly initialized Portuguese encoder $\phi_{E_{\mathcal{Y}}}$ is trained from scratch, which has the same architecture as the English encoder T^X_{en}. We could obtain the Portuguese encoder output T^Y_{en} 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. Smooth L1 loss is leveraged to optimize the encoder-level KD:

$$\mathcal{L}_{kd} = \text{smoothL1}(T^X_{en}, T^Y_{en}) \tag{8}$$

where smoothL1(..) denotes the smoothL1 loss.

The Encoder-level KD could also be extended to the middle sub-layers of the bidirectional transformer encoder.

(2) Vocab-level KD: as shown in Fig. 4 (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. The vocab-level KD process is fulfilled by minimizing the smooth L1 loss between the English embedding H^X and the Portuguese embedding H^Y:

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

Specially, during the KD process, a crucial problem is that the parallel sequence lengths of different language transcripts are diverse. To overcome this problem, we adopt an adaptive pooling[2] mechanism to transform the English word feature sequence to the same length as the Portuguese one.

Table 1: Statistics of the data partition, where the numbers represent the number of videos. The data in How2-MCLS and How2-MS datasets do not overlap, and the Portuguese videos in How2-MCLS also have corresponding Portuguese and English transcripts.

Partition	Training	Validation	Test
How2-MCLS: Video+Pt2En	13,167	150	127
How2-MS: Video+En2En	59,539	-	-

2.5 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. 2 (c), we could replace the English encoder $\phi_{D_{\mathcal{X}}}$ (vocab $\phi_{V_{\mathcal{X}}}$) with the Portuguese encoder $\phi_{E_{\mathcal{Y}}}$ (vocab $\phi_{V_{\mathcal{Y}}}$) to form a new VDF model $\phi_{\mathcal{Y}}$. Finally, we utilize the new VDF model $\phi_{\mathcal{Y}}$ to read 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:

$$\hat{S} = \phi_{\mathcal{Y}}(V, Y) \tag{10}$$

$$\mathcal{L}_{ft} = -\sum_{t=l}^{L} \log P\left(S_t \mid \hat{S}_{<t}, V, Y\right) \tag{11}$$

3 Experiments

3.1 Dataset Construction

We conduct the experiments by reorganizing the How2 dataset (Sanabria et al., 2018). It is a large-scale open-domain instructional video dataset, which includes three sub-task data: multimodal

[2]please refer to torch.nn.AdaptiveAvgPool1d.

Table 2: Experiment results in the general MCLS. Pipe-[MODEL] means that first translate the source transcript to the other language (Pt2En) in the test set, then use the monolingual multimodal summarization [MODEL] to generate summaries. We use the Google translation system to strengthen Pipe-[MODEL] baselines.

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
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	41.35	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**

machine translation (MMT), multimodal speech recognition (MSR), and multimodal summarization (MS). The MS data includes 2,000h of videos accompanied by English transcripts and summaries. Besides, MMT data includes 300h 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 300h videos and bilingual transcripts, combined with the summaries provided by MS, as **How2-MCLS** dataset for the general scenario; 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 for the MS-augmented scenario. The dataset statistics are shown in Table 1.

3.2 Implementation Details

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 for 50 epochs with a batch size of 8 on 1 NVIDIA Tesla v100 GPU, and we adopt cross-entropy loss and 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. (2019), 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.

3.3 Baseline Models

As MCLS is a newly-proposed task, we construct some recent multimodal summarization baselines of single or multiple modalities for comparison: **S2S** (Luong et al., 2015): a standard sequence-to-sequence RNN with attention mechanism. **NCLS** (Zhu et al., 2019): a transformer-based encoder-decoder model for cross-lingual summarization. **VideoRNN** (Palaskar et al., 2019): a sequence-to-sequence RNN model that receives video features to generate summaries. **MT** (Zhou et al., 2018): a transformer-based encoder-decoder architecture transforming video sequence features to captions. **HA** (Palaskar et al., 2019): a multisource sequence-to-sequence model with a hierarchical attention to combine video and text modalities. **MFFG** (Liu et al., 2022): a multistage fusion network with the forget gate module for multimodal summarization. These baselines are compared with our proposed method: **VDF**: video-guided dual fusion network; **VDF-TS-V(E)**: **T**riple-**S**tage training framework utilizing VDF as the backbone and **V**ocab (**E**ncoder)-level knowledge distillation.

3.4 Overall performance

BLEU (1,2,3,4) (Papineni et al., 2002), ROUGE (1,2,L) (Lin, 2004), METEOR (Banerjee and Lavie, 2005) and CIDEr (Vedantam et al., 2015) are adopted as the evaluation metrics to comprehensively analyze the model performance. In Table 2, we simulate the general MCLS scenario by only utilizing the How2-MCLS dataset. It can be observed that: (1) Multimodal models typically perform better than unimodal models, illustrating the importance of multi-source information to promote summary generation. (2) The pipeline methods of first translating and then monolingual

Table 3: Experiment results in the MS-augmented MCLS, where sufficient English video-summary data in the How2-MS dataset are utilized to assist the models. P:pre-training; F:fine-tuning; T:translation; M:mix-training. [MODEL]-TS-V(E) denotes **t**riple-**s**tage training framework with [MODEL] backbone and **v**ocab(**e**ncoder)-level knowledge distillation.

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
VDF (T+P+F)	49.09	38.30	32.41	28.44	51.37	32.23	46.33	22.63	1.774
MFFG-TS-V	47.69	36.62	30.68	26.65	50.70	30.50	45.47	22.16	1.726
MFFG-TS-E	46.97	36.11	30.31	26.52	49.74	29.60	44.15	21.69	1.753
VDF-TS-V	49.37	38.82	33.13	29.26	51.75	33.14	46.95	22.88	1.875
VDF-TS-E	**50.01**	**39.26**	**33.47**	**29.50**	**52.16**	**33.31**	**47.18**	**23.19**	**1.910**

Table 4: 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

summarizing do not perform well. A crucial reason is the error propagation in translation. (3) Our VDF model outperforms the strong multimodal baseline models and achieves the state-of-the-art performance, which shows the effectiveness of the proposed dual fusion strategies.

In Table 3, we compare the VDF model under the proposed triple-stage training framework (i.e., VDF-TS-E and VDF-TS-V) with the recent state-of-the-art model MFFG in the MS-augmented MCLS scenario. In the experiment, we construct the following three kinds of comparison variations: a) first pre-train on How2-MS, then fine-tune on How2-MCLS (P+F); b) mix How2-MS and How2-MCLS for training (M); c) train an English to Portuguese translation model based on BART (Lewis et al., 2020) with the bilingual transcripts in How2-MCLS, and then conduct a) with translated PT transcripts (T+P+F). From the experiment results, we could observe that in the cases of the same experimental settings, VDF-based models consistently outperforms MFFG-based models; and VDF-TS-E and VDF-TS-V outperform all the compared models, demonstrating the superiority of our proposed method. Moreover, we also conduct the proposed triple-stage training framework with encoder/vocab-level knowledge distillation strategy on MFFG, i.e., constructing MFFG-TS-E and MFFG-TS-V. Experiment results show that with the same backbone architecture, MFFG-TS-E and MFFG-TS-V surpass the other MFFG-based models, which illustrates the effectiveness of the triple-stage training framework.

3.5 Ablation Analysis

We construct ablation experiments to demonstrate the validity 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 VDF-TS-V and VDF-TS-E frameworks, 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 4, and we could draw the following conclusions: (1) Eliminating either FF-Enc or CF-Dec components causes performance degradation of the VDF model, revealing that the two fusion structures are effective. (2) The model performance dramatically decreases after removing the video input and video-related structures, which demonstrates the importance of multimodal information. (3) Knowledge distillation mechanism brings 1.77 absolute ROUGE-L points promotion for VDF-TS-E. We believe the reason is that the triple-stage training framework could effectively transfer the prior knowledge from the well-informed English teacher to the Portuguese student, thus achieving

Table 5: Knowledge distillation on different encoder sub-layers.

Method	ROUGE-1	ROUGE-2	ROUGE-L
VDF-TS-E w/o KD	50.46	31.16	45.41
VDF-TS-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

better performances.

3.6 Knowledge Distillation Analysis

We explore the impact of the triple-stage training framework 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. From the experiment results in Table 5, we could observe that: (1) No matter the knowledge distillation is performed in any encoder sub-layer, the proposed methods outperform those without distillation, which illustrates the effectiveness of the knowledge distillation module. (2) The best experiment results are obtained in the middle encoder layer (layer=2). Such results even exceed the performance reported in the main experiment.

3.7 Triple-stage Training in Low-resource Scenario

To investigate the model performance in lower resource MCLS scenarios, we conduct experiments by reducing the cross-lingual video-summary samples. To be specific, in the first phase of the proposed triple-stage training framework, 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. 5. We could observe that using as few as 1k samples in the How2-MCLS dataset, VDF-TS-V and VDF-TS-E achieve comparable or better performance than the best unimodal models (e.g., NCLS and MT), which utilize the 13k samples for training. Meanwhile, with only leveraging 3k How2-MCLS data, VDF-TS-V and VDF-TS-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 framework in the low-resource MCLS situation.

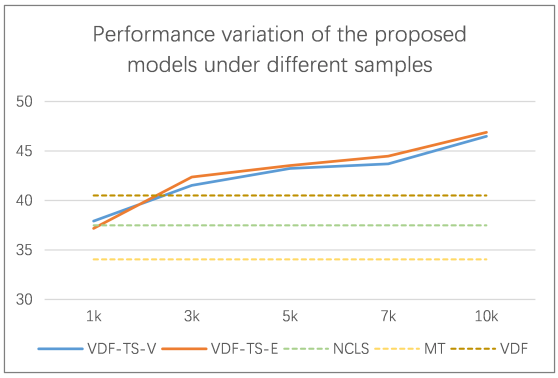

Figure 5: ROUGE-L performance variations of the proposed models under different Portuguese video-English summary training samples. The dotted lines in the figure illustrate the models trained on the whole How2-MCLS dataset, and the solid lines are the results of the proposed triple-stage model with different sample sizes.

4 Related Work

To allow non-native speakers holistically understand the videos in other languages, we propose the Multimodal Cross-Lingual Summarization for videos (MCLS) task. As MCLS is a newly proposed task, there is no existing work. The most relevant research area includes: multimodal summarization (MS) (Khullar and Arora, 2020; Zhu et al., 2018; Yu et al., 2021; Liu et al., 2021; Zhang et al., 2022b,a) and textual cross-lingual summarization (Cao et al., 2020; Zhu et al., 2020; Ouyang et al., 2019; Xu et al., 2020). Between the two, MS seeks to compresses multimedia documents. Sanabria et al. (2018) first released the How2 dataset for multimodal abstractive summarization for open-domain videos, which provides multisource information, including videos, audios, text transcriptions and human-generated summaries. With the rise of sequence-to-sequence learning (Sutskever et al., 2014; See et al., 2017; Wu et al., 2020), Khullar and Arora (2020) utilized hierarchical attention to fuse text, audio and video modalities. Liu et al. (2020) proposed a multistage forget gate to resist the flow of multimodal noise. Meanwhile, textual cross-lingual summarization is the task of generating a target-language summary from the given documents in the source language, and some textual cross-lingual methods are reported to perform well (Cao et al., 2020; Zhu et al., 2020). However, existing summarization approaches could not consider multilingualism and multimodality simultaneously. Moreover, multimodal cross-lingual summarization task is likely to meet the data-insufficient

problem. Despite some textual cross-lingual summarization approaches utilising external toolkits (Jiang et al., 2022) or pre-trained language models (PLMs) (Xu et al., 2020) to overcome such a low-resource dilemma, they can not be directly applied to the newly proposed MCLS task. In this work, we propose the VDF model and the triple-stage training framework to address the above problems.

5 Conclusions

In this work, we propose a new task: 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 video-guided dual fusion network (VDF) to integrate multimodal and cross-language information for summary generation. In addition, we introduce a triple-stage training framework to enhance models in MCLS with monolingual multimodal summarization data. Experiment results illustrate the effectiveness of the VDF model and the triple-stage training framework. Specially, in the low-resource MCLS scenario, the proposed methods achieve comparable or better performance than those baseline models trained with the full amount of data.

Limitations

We have to admit that our work has the following limitations:

1) In our proposed triple-stage training framework, the second stage training relies on a small amount of bilingual parallel corpus for knowledge distillation.

2) For a fair comparison, we adopt the bilingual transcript provided by How2 dataset as the parallel corpus in the second stage of triple-stage training framework. In fact, this part of the corpus can be replaced by parallel corpus from other source for knowledge distillation, which makes our method more practicable. Although the effectiveness of the above modules has been verified by using the How2 dataset, we did not conduct extended experiments which use more external parallel corpora for knowledge distillation of second training stage to further verify the applicability of our model. We plan to conduct these extended experiments in the future.

6 Acknowledgements

We thank the support of Natural Science Foundation of China under Grant 62206267.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Yue Cao, Hui Liu, and Xiaojun Wan. 2020. Jointly learning to align and summarize for neural cross-lingual summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6220–6231. Association for Computational Linguistics.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555.

Shuyu Jiang, Dengbiao Tu, Xingshu Chen, Rui Tang, Wenxian Wang, and Haizhou Wang. 2022. Cluegraphsum: Let key clues guide the cross-lingual abstractive summarization. *CoRR*, abs/2203.02797.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Aman Khullar and Udit Arora. 2020. MAST: multimodal abstractive summarization with trimodal hierarchical attention. *CoRR*, abs/2010.08021.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Nayu Liu, Xian Sun, Hongfeng Yu, Fanglong Yao, Guangluan Xu, and Kun Fu. 2022. Abstractive summarization for video: A revisit in multistage fusion network with forget gate. *IEEE Transactions on Multimedia*.

Nayu Liu, Xian Sun, Hongfeng Yu, Wenkai Zhang, and Guangluan Xu. 2020. Multistage fusion with forget gate for multimodal summarization in open-domain videos. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1834–1845. Association for Computational Linguistics.

Nayu Liu, Xian Sun, Hongfeng Yu, Wenkai Zhang, and Guangluan Xu. 2021. D-mmt: A concise decoder-only multi-modal transformer for abstractive summarization in videos. *Neurocomputing*, 456:179–189.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1412–1421.

Jessica Ouyang, Boya Song, and Kathleen McKeown. 2019. A robust abstractive system for cross-lingual summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2025–2031.

Shruti Palaskar, Jindrich Libovický, Spandana Gella, and Florian Metze. 2019. Multimodal abstractive summarization for how2 videos. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6587–6596. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the annual meeting on association for computational linguistics (ACL)*, pages 311–318. Association for Computational Linguistics.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A large-scale dataset for multimodal language understanding.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Xindi Shang, Zehuan Yuan, Anran Wang, and Changhu Wang. 2021. Multimodal video summarization via time-aware transformers. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 1756–1765. ACM.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, pages 6000–6010.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4566–4575.

Yiquan Wu, Kun Kuang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Jun Xiao, Yueting Zhuang, Luo Si, and Fei Wu. 2020. De-biased court's view generation with causality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 763–780.

Ruochen Xu, Chenguang Zhu, Yu Shi, Michael Zeng, and Xuedong Huang. 2020. Mixed-lingual pretraining for cross-lingual summarization. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020*, pages 536–541. Association for Computational Linguistics.

Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021. Vision guided generative pre-trained language models for multimodal abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3995–4007.

Litian Zhang, Xiaoming Zhang, and Junshu Pan. 2022a. Hierarchical cross-modality semantic correlation learning model for multimodal summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11676–11684.

Zijian Zhang, Chang Shu, Youxin Chen, Jing Xiao, Qian Zhang, and Lu Zheng. 2022b. Icaf: Iterative contrastive alignment framework for multimodal abstractive summarization. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. 2018. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8739–8748.

Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. Msmo: Multimodal summarization with multimodal output. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4154–4164.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. Ncls: Neural cross-lingual summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3054–3064.

Junnan Zhu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2020. Attend, translate and summarize: An efficient method for neural cross-lingual summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1309–1321. Association for Computational Linguistics.