Learning Summary-Worthy Visual Representation for Abstractive Summarization in Video

Zenan Xu[1*] , **Xiaojun Meng**[2] , **Yasheng Wang**[2] , **Qinliang Su**[1,4†] ,
Zexuan Qiu[3] , **Xin Jiang**[2] and **Qun Liu**[2]
[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2]Noah's Ark Lab, Huawei Technologie
[3]The Chinese University of Hong Kong, Hong Kong SAR
[4]Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou, China
{xuzn@mail2, suqliang@mail}.sysu.edu.cn, qzexuan@link.cuhk.edu.hk,
{xiaojun.meng, wangyasheng, Jiang.Xin, qun.liu}@huawei.com

Abstract

Multimodal abstractive summarization for videos (MAS) requires generating a concise textual summary to describe the highlights of a video according to multimodal resources, in our case, the video content and its transcript. Inspired by the success of the large-scale generative pre-trained language model (GPLM) in generating high-quality textual content (e.g., summary), recent MAS methods have proposed to adapt the GPLM to this task by equipping it with the visual information, which is often obtained through a general-purpose visual feature extractor. However, the generally extracted visual features may overlook some summary-worthy visual information, which impedes model performance. In this work, we propose a novel approach to learning the summary-worthy visual representation that facilitates abstractive summarization. Our method exploits the summary-worthy information from both the cross-modal transcript data and the knowledge that distills from the pseudo summary. Extensive experiments on three public multimodal datasets show that our method outperforms all competing baselines. Furthermore, with the advantages of summary-worthy visual information, our model can have a significant improvement on small datasets or even datasets with limited training data.

1 Introduction

With the increasing popularity of video in user-generated content in recent years [Kim *et al.*, 2021; Cherian *et al.*, 2022], a large number of open-domain videos have been posted on the Web (e.g., YouTube). Usually, many of the videos are not accompanied by a briefly introduction to reflect the corresponding salient information, which prevents users from quickly finding their interested videos unless taking time to peruse each video. Obviously, in this scenario, it would be valuable to develop an automatic abstractive summarization model [Libovickỳ *et al.*, 2018] that detects the highlights of each video and then generates a short textual description.

Transcript: so, now we 're going to talk about the **backhand**. basically, the same a grip. sometimes i feel like i want to adjust it [...] you 're hanging your arm, trying to hang your arm almost perpendicular to the ground. almost like a tennis **stroke**. [...] and just follow through, you can come through a little high, but basically, you want to get your arms back into the ready position. we 'll talk about that next.

Reference Summary: learn how to do the **backhand stroke** in ping pong in this free video lesson on **ping pong** .

Figure 1: An example of multimodal abstractive summarization. The unimportant textual content is omitted and replaced by [...]. It can be seen that some critical information in the reference is either highlighted in the transcript (e.g., **backhand stroke**) or only available from the video (e.g., **ping pong**).

As illustrated in Figure 1, the task of multimodal abstractive summarization (MAS) aims to generate a concise textual summary according to multimodal resources, i.e., the video content and its transcript [Sanabria *et al.*, 2018]. This task is challenging since both visual and textual modalities are complementary to each other, and thus, how to efficiently combine the multimodal information is the key to this task. To leverage information from both modalities, [Palaskar *et al.*, 2019] employed separate encoders on visual and textual data, which was followed by a joint decoder with an attention mechanism to capture the intrinsic connection between the two modalities. Later, MFFG [Liu *et al.*, 2020] and SFFG [Liu *et al.*, 2022] brought the multimodal interaction into the encoder to obtain the fine-grained correlation between multimodal inputs to exploit the complementary information of each modality and achieved promising results.

Recently, inspired by the success of the large-scale generative pre-trained language model (GPLM) [Lewis *et al.*, 2019; Raffel *et al.*, 2020] on generating high-quality textual content (e.g., summary), researchers start to apply it to the MAS

[*] Work is done during internship at Noah's Ark Lab.

[†] Corresponding author.

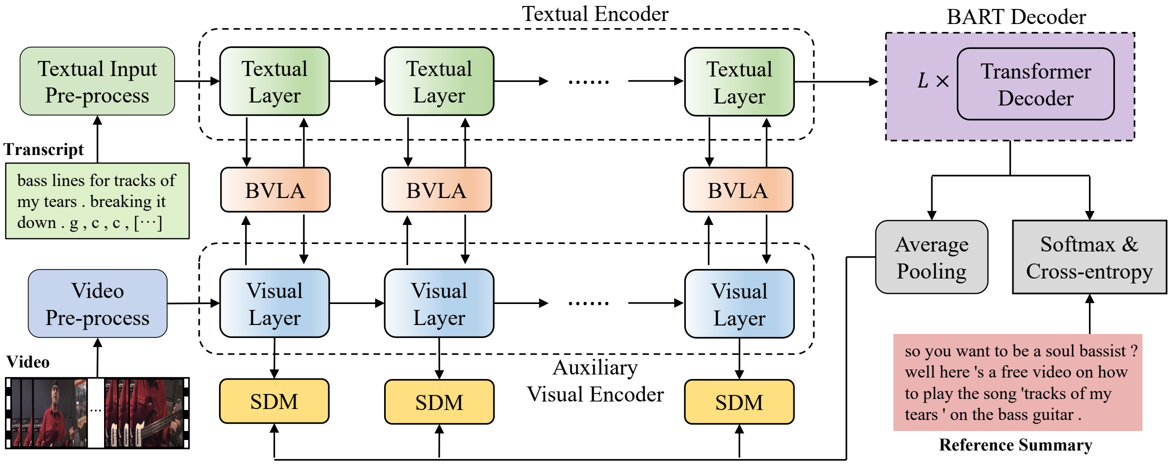


Figure 2: An overview of our proposed *SWVR* (Summary-Worthy Visual Representation) method. The Bi-directional Visual-Language Attention mechanism (BVLA) and Self Distillation Mechanism (SDM) are introduced to encourage the visual encoder to exploit the summary-wothy information.

task. To do this, VG-GPLM [Yu *et al.*, 2021a] first extracts the video representation with the pre-trained visual feature extractor. Then, to allow the cross-modal interaction, the obtained visual feature of video is injected into each encoder layer of the GPLM with an attention-based text-vision fusion module. This attention module is designed to select the relevant visual information (key & value) based on the textual features (query). VG-GPLM has demonstrated the potentials of injecting visual representation into language models to improve the MAS performance.

However, existing methods mainly use a general-purpose visual feature extractor, which may potentially overlook some summary-worthy visual information, including the essentials that are merely available in the video (e.g., the ‘`ping pong`’ in Figure 1 only appears in video frames), and thus impede the quality of the generated summary. It is also possible for the existing methods to generate the novel concepts like ‘`ping pong`’ if sufficient labelled data is provided to well-train the general-purpose extractor to identify task-specific objects. Unfortunately, it may lead to the contradiction between the data hungriness issue of the extractor and the scarcity of annotated summarization data.

To address the above issue, we aim to design a visual encoder that is aware of summary-worthy information. We present a novel method, named *SWVR*, to learn **S**ummary-**W**orthy **V**isual **R**epresentation for the MAS task. We use a **B**i-directional **V**isual-**L**anguage **A**ttention mechanism (BVLA) to encourage the visual encoder to exploit the summary-wothy information from the textual data, i.e., transcripts. Afterward, we further introduce a **S**elf **D**istillation **M**echanism (SDM) that takes the generated pseudo summary as the teacher to guide the learning of the visual encoder. This self-distillation encourages the visual encoder to have the ability of capturing and aligning to summary-worthy knowledge that appears in the generated summary after the decoder. Given the surprising and strong text generation ability of GPLM, such distillation helps the visual encoder pay more attention on corresponding key concepts in video frames. Since a sub-optimal placement may impede the model's performance, we enumerate almost all possible ways to combine and evaluate BVLA and SDM modules in the visual encoder. Experimental results on three public datasets show that our method outperforms all competing baselines, especially on the dataset with a relatively small size. Further studies demonstrate the effectiveness of each component, and suggest that the learned summary-worthy feature can help the model identify valuable information, thus benefiting the MAS task.

2 Related Work

Text-based Abstractive Summarization: Given a long article, abstractive text summarization aims to generate a brief summary that describes the article's most salient content. Early studies were based on statistical or linguistic rules, including extractive and compression [Knight and Marcu, 2002; Clarke and Lapata, 2010], templates [Zhou and Hovy, 2004], and statistics [Banko *et al.*, 2000]. Later, the availability of large-scale summarization corpora has promoted the development of various neural networks, among which the representative Seq2Seq model [Sutskever *et al.*, 2014] and the attention mechanism have greatly advanced the quality of summaries [Paulus *et al.*, 2018; Wang *et al.*, 2019; Zhang *et al.*, 2020; Yu *et al.*, 2021b]. Recently, in light of the powerful generative abilities, the large-scale pre-trained language models have led the mainstream in this field [Lewis *et al.*, 2019; Raffel *et al.*, 2020; Zhang *et al.*, 2020; Qi *et al.*, 2020].

Video-based Abstractive Summarization: Multimodal summarization has been developed for decades [Erol *et al.*, 2003; Tjondronegoro *et al.*, 2011; Evangelopoulos *et al.*, 2013; Shah *et al.*, 2016; Zhang *et al.*, 2022]. The key method behind it, i.e., multimodal learning, has recently attracted a number of researchers' interest, while in fact little attention has been paid to the video content based summarization. Previous methods mainly focus on a simple situation where the video data contains synchronized signals, e.g., synchronized voice and captions. To tackle the video-based summarization in a more general and asynchronous scenario, [Li *et al.*, 2017] collected a multimodal dataset containing 500 videos of English news articles with human-generated reference summaries. Later, to better promote the development of the MAS for videos, [Sanabria *et al.*, 2018] introduced a large-scale human-annotated video dataset named How2, which contains videos of 2000 hours, and each video is annotated with a short summary. Thanks to the How2 dataset, of which the advent has accelerated the development of MAS methods using neural networks, e.g., the hierarchical attention in [Palaskar *et al.*, 2019], the forget gate mechanism in [Liu *et al.*, 2020], and the multi-stage fusion network in [Liu *et al.*, 2022]. To further leverage the GPLM's generation ability, VG-GPLM [Yu *et al.*, 2021a] studies multiple fusion methods that inject the visual information into GPLMs to improve the MAS for videos. However, VG-GPLM obtains the visual feature via a general-purpose visual encoder, which likely ignores task-specific visual clues that are valuable to summarization. In contrast, our method bridges this gap by learning and then injecting the summary-worthy visual feature into the GPLMs.

3 Methodology

3.1 Problem Definition

Given a video V and its associated textual transcript T, our multimodal summarization system is required to summarize the video by a generated concise summary S with maximum probability $p(S|V,T;\theta)$, where θ stands for the model parameters. It is worth knowing that both the transcript T and generated summary S are in the form of a sequence of words.

For the text-based abstractive summarization task, a prevailing way is to adopt the powerful GPLM, whose ability of generating high-quality texts has been widely demonstrated [Qi *et al.*, 2020; Zhang *et al.*, 2020]. When in the case of multimodal input like a video, a natural idea is to inject the visual features into the encoder of GPLM and then utilize its superior generation to obtain a better textual summary, which is how VG-GPLM [Yu *et al.*, 2021a] works. In the following, we present the proposed *SWVR* model in detail, and the overall framework is shown in Figure 2.

3.2 Video Pre-processing

Given a video, we follow previous work [Yu *et al.*, 2021a] to first extract the visual features with 2048 dimension for every 16 non-overlapping frames through a 3D ResNeXt-101 model [Hara *et al.*, 2018], which is pre-trained on the Kinetics dataset [Kay *et al.*, 2017]. Then a linear layer is adopted to project the visual features into d dimension vector space and obtain $\boldsymbol{Z}_v \in \mathbb{R}^{n_v \times d}$, where n_v is the number of visual tokens.

3.3 Method Overview

As discussed above, we employ the GPLM as our backbone. Specifically, we construct our method with the sequence-to-sequence BART model [Lewis *et al.*, 2020], which consists of a textual encoder and a left-to-right decoder designed to generate the textual summary. Given the transcript as input, the textual encoder first tokenizes it and then embeds it into the textual features $\boldsymbol{Z}_t \in \mathbb{R}^{n_t \times d}$, which will be fed into the transformer [Vaswani *et al.*, 2017] encoder layers to exploit the contextual representation as:

$$\boldsymbol{Z}_t^{l'} = LN(MultiAttn(\boldsymbol{Z}_t^{l-1}) + \boldsymbol{Z}_t^{l-1}), \tag{1}$$

$$\boldsymbol{Z}_t^{l} = LN(FFN(\boldsymbol{Z}_t^{l'}) + \boldsymbol{Z}_t^{l'}), \tag{2}$$

where LN denotes the layer normalization [Ba *et al.*, 2016], $l \in [1, L]$ denotes the l-th textual encoder layer, $\boldsymbol{Z}_t^0$ is initialized with $\boldsymbol{Z}_t$, and $MultiAttn$ and FFN denotes two sub-layers, i.e., the multi-headed self-attention mechanism and a feed-forward network, respectively[1].

To inject the visual information into each layer of the textual encoder, we add a third sub-layer into each textual encoder layer. Specifically, assume the visual feature that is to be injected into the l-th textual encoder layer is $\boldsymbol{Z}_v^l$, then this sub-layer operates as:

$$\tilde{\boldsymbol{Z}}_t^{l} = LN(VLA(\boldsymbol{Z}_v^l, \boldsymbol{Z}_t^l) \cdot \boldsymbol{W}_t^l + \boldsymbol{Z}_t^l), \tag{3}$$

where $\boldsymbol{W}_t^l \in \mathbb{R}^{d \times d}$ is the model weight, and $VLA(\boldsymbol{Z}_v^l, \boldsymbol{Z}_t^l)$ is a unidirectional visual-language attention function that extracts the relevant information from visual feature $\boldsymbol{Z}_v^l$ according to the textual feature $\boldsymbol{Z}_t^l$, which will be described later. It can be seen from (3) that $\boldsymbol{Z}_v^l$ takes a great impact on our model. Therefore, our paper focuses on how to obtain a high-quality collection of visual feature, i.e., $\boldsymbol{C}_v = \{\boldsymbol{Z}_v^1, \cdots, \boldsymbol{Z}_v^L\}$, where each $\boldsymbol{Z}_v^l$ can be injected into the l-th textual encoder layer.

Afterward, the output $\tilde{\boldsymbol{Z}}_t^L$ from the last layer of textual encoder will be provided to the decoder to generate a sequence $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{n_w})$ of vectors one element at a time. Since the decoder is auto-regressive, at each step, the generated vector $\boldsymbol{w}_i$ is mapped to a new vector of vocabulary size, followed by a softmax function to output the summary word distribution $\hat{\boldsymbol{y}}_i$. The negative log-likelihood between the generated summary and the ground-truth summary is used to calculate the loss as:

$$\mathcal{L}_{Abs} = -\sum_{j=1}^{n_w} \boldsymbol{y}_i \log p(\hat{\boldsymbol{y}}_i). \tag{4}$$

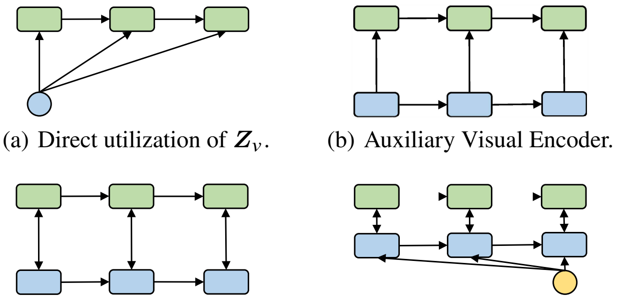
3.4 Learning Summary-Worthy Visual Representation

A simple way to generate the collection $\boldsymbol{C}_v$ is that we fill $\boldsymbol{C}_v$ with the obtained $\boldsymbol{Z}_v$ after video pre-processing, namely, $\boldsymbol{C}_v = \{\boldsymbol{Z}_v^l = \boldsymbol{Z}_v, \forall l = 1, \cdots, L\}$. In this way, the cross-modal interaction can be simply summarized as Figure 3(a).

Auxiliary Visual Encoder

However, the direct utilization of $\boldsymbol{Z}_v$ may suffer from the below shortcomings: (*i*) The temporal information of a video is ignored, which limits the expressiveness of visual feature; (*ii*) Since recent study [Raghu *et al.*, 2021] finds that different encoder layers exploit different-level semantic, injecting the same visual feature into different textual encoder layers may potentially restrict the model to learn more meaningful cross-modal interactions.

[1] Additionally, there are residual operations inside each layer, which we omit here for brevity. For the interested readers, we kindly refer to [Vaswani *et al.*, 2017] for more description.



(a) Direct utilization of Z_v. (b) Auxiliary Visual Encoder.

(c) Learning SWR from Transcript. (d) Learning SWR from Summary.

Figure 3: The simple illustration of different types of visual information injected into the textual encoder layers. SWR stands for summary-worthy representation. The textual and visual information are represented in **green** and **blue** color, respectively. The **yellow** color stands for the feature map of generated summary.

To address the above issues, also inspired by the recent success of the two-tower architecture [Radford *et al.*, 2021; Jia *et al.*, 2021], we introduce an auxiliary visual encoder that takes the video feature as input and learns visual information at different levels of semantics. Specifically, the trainable positional encodings are first added to the Z_v to preserve the temporal information. Afterward, this Z_v can be used directly as input to vanilla transformer-based visual encoder of L layers. Formally, given the input Z_v, the output of each visual layer is computed as:

$$Z_v^{l'} = LN(MultiAttn(Z_v^{l-1}) + Z_v^{l-1}), \tag{5}$$

$$Z_v^{l} = LN(FFN(Z_v^{l'}) + Z_v^{l'}), \tag{6}$$

where Z_v^0 is initialized with Z_v. We collect the output of each visual encoder layer and obtain a new visual collection $C_v = \{Z_v^1, \cdots, Z_v^L\}$, as shown in Figure 3(b). In this way, we inject the visual information into the textual feature from the same 'height' (e.g., both from the l-th layer), which could benefit the cross-modal correlation at different levels of semantics.

Summary-Worthy Information from Transcript

It can be observed that this introduced auxiliary visual encoder is still a general-purpose visual feature extractor. Since not all the information from the visual modality is valuable for summarization [Liu *et al.*, 2020], the noise and redundancy in the generally learned visual feature makes it less effective for the model to summarize the highlights. We are thus motivated to propose to learn the summary-worthy visual representation.

Since the summary-worthy information from each modality may be complementary to the other, carrying this observation, we propose to exploit the summary-worthy information from the transcript, which may be complementary to and benefit learning the visual feature. Therefore, we use the multi-head cross-attention to inject the summary-worthy information from the transcript into the visual representation. Technically, we extend the unidirectional $VLA(\cdot,\cdot)$ function (which is employed in (3)) into bi-directional visual-language attention function $BVLA(\cdot,\cdot)$. Given textual feature $Z_t^l \in \mathbb{R}^{n_t \times d}$ and visual representation $Z_v^l \in \mathbb{R}^{n_v \times d}$, the $BVLA(\cdot,\cdot)$ will function as:

$$[Z_{t\to v}^l, Z_{v\to t}^l] = BVLA(Z_t^l, Z_v^l), \tag{7}$$

where $Z_{t\to v}^l \in \mathbb{R}^{n_v \times d}$ and $Z_{v\to t}^l \in \mathbb{R}^{n_t \times d}$ denote the information of transcript that is relevant to video or vice versa, respectively. Let us take $Z_{t\to v}^l$ as an example to elaborate. Particularly, the $Z_{t\to v}^l$ is calculated via a cross-modal multi-head attention ($CrossMultiAttn$) as:

$$Q_v = Z_v^l W_q^l, \;\; K_t = Z_t^l W_k^l, \;\; V_t = Z_t^l W_v^l, \tag{8}$$

$$Z_{t\to v}^l = CrossMultiAttn(K_t, Q_v, V_t), \tag{9}$$

where W_q^l, W_k^l and $W_v \in \mathbb{R}^{d\times d}$ are model weights. After obtaining the $Z_{t\to v}^l$, we inject it into the visual feature as

$$\tilde{Z}_v^l = LN(Z_{t\to v}^l \cdot W'^l_v + Z_v^l), \tag{10}$$

where $W'^l_v \in \mathbb{R}^{d\times d}$ is the model weight. Similarly, the output of each visual encoder layer can be collected to form $C_v = \{\tilde{Z}_v^1, \cdots, \tilde{Z}_v^L\}$, as like Figure 3(c).

Summary-Worthy Information from Summary

Since some vital summary concepts are only available in the video content rather than transcripts (e.g., the '`ping pong`' in Figure 1), we also expect the visual encoder could have the ability to identify the novel summary-worthy visual feature that might not appear in transcripts. We are thus motivated to bridge the visual feature learning and the generated summary in a more direct manner, thus further enhancing our visual encoder.

One potential way of benefiting our visual encoder is to distil the task-specific knowledge from other large-scale summarization models. However, most of the available ones are text-only methods, of which the knowledge is not suitable for our visual encoder. To this end, instead of the traditional knowledge distillation, we take advantage of self distillation [Zhang *et al.*, 2019], which has been proven to be an effective way to distil knowledge within the network itself. The basic idea is that we take the output of the decoder as the pseudo summary (i.e., hints) to teach the learning process of the visual encoder.

Technically, we apply the average pooling on the output vector sequence $(w_1, \ldots, w_{n_w})$ from the last layer of the decoder to obtain $w \in \mathbb{R}^d$, which can be regarded as the global representative of the generated pseudo summary. Then we distil this summary-worthy knowledge into our visual encoder layer via a self distillation mechanism (SDM) as

$$\mathcal{L}_{Sd} = \lambda \cdot \|w - Linear(avg(\tilde{Z}_v^l))\|, \tag{11}$$

where λ is the hyper-parameters, $avg(\cdot)$ is the average pooling layer. As seen from (11), this knowledge distillation works by decreasing the distance between the feature maps of the pseudo summary from the decoder and of the visual feature from the visual encoder. Since the feature maps are from two modalities with different dimensions, an extra linear projector is added to project the visual one to the same dimension as the other. Finally, the training loss $\mathcal{L}$ of our method is a sum of the objectives as:

$$\mathcal{L} = \mathcal{L}_{Abs} + \mathcal{L}_{Sd}. \tag{12}$$

Dataset	Train	Dev	Test
How2	68336	2520	2127
How2-300	13167	150	127
MM-AVS	836	104	105

Table 1: The statistics of the three datasets.

During the training, as the visual feature in our SDM can gradually fit the feature map of the decoder output in a global representative way, the inexplicit knowledge and novel concepts that have not appeared in transcripts are injected into learning the visual feature, and thus we achieve a new summary-worthy collection $\boldsymbol{C}_v = \{\hat{\boldsymbol{Z}}_v^1, \cdots, \hat{\boldsymbol{Z}}_v^L\}$, see Figure 3(d).

3.5 Implementation Details

The BART-base model is adopted as the backbone of our model, in which $L = 6$ for both the encoder and decoder. For the introduced auxiliary visual encoder, we use a 6-layer encoder with 8 attention heads and a 768 feed-forward dimension. Following previous work [Yu *et al.*, 2021a], we set the max length of the generated summary to be 64 tokens; the decoding process can be stopped early if an End-of-Sequence (EOS) token is emitted. The Adam [Kingma and Ba, 2014] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of $1e^{-5}$ is employed as the optimizer.

4 Experiments

4.1 Experimental Setups

Datasets and Evaluation Metrics: We evaluate the proposed *SWVR* on three public datasets, including How2, How2-300 [Sanabria *et al.*, 2018], and MM-AVS [Fu *et al.*, 2021] dataset. The statistic of datasets is shown in Table 1. We follow [Yu *et al.*, 2021a] to adopt the ROUGE-{1,2,L} [Lin and Hovy, 2003], BLEU-{1,2,3,4} [Papineni *et al.*, 2002], METEOR [Denkowski and Lavie, 2011], CIDEr [Vedantam *et al.*, 2015], and Content F1 [Palaskar *et al.*, 2019] as the evaluation criteria.

Bacelines: We compare *SWVR* with the following two groups of baselines. *1) Methods using transcript only:* In this group of baselines, we pick the existing text-only summarization methods, including S2S [Luong *et al.*, 2015], PG [See *et al.*, 2017], TF [Vaswani *et al.*, 2017], T5 [Raffel *et al.*, 2020], and BART [Lewis *et al.*, 2020]. *2) Methods using both transcript and video:* We consider strong MAS baselines including HA(RNN/Transformer) [Palaskar *et al.*, 2019], MFFG(RNN/Transformer) [Liu *et al.*, 2020], VG-GPLM(T5/BART) [Yu *et al.*, 2021a], and SFFG [Liu *et al.*, 2022].

4.2 Experimental Results

The results of How2 datasets are shown in Table 2. As observed, our proposed model significantly outperforms all previous methods, and we take an average improvement of 5.72% over all the criterion compared with the strongest baseline VG-BART. It confirms the effectiveness of our proposed summary-worthy mechanism. We can also see that the performance of abstractive summarization with the help of both transcripts and video content significantly outperform transcript-only summarization models, demonstrating that visual features contain valuable information that is complementary to the transcript.

As for the How2-300 dataset, it can be observed from Table 3 that our model still performs the best. An interesting found is that, compared with the VG-BART model, the average improvement 4.9% (over R1, R2, and RL) on How2-300 is larger than that of 2.2% in the How2 dataset, namely, our model can work even better on small datasets like How2-300. We conjecture that it is because the general-purpose visual feature extractors require more training data to distinguish the task-specific clues that are valuable for the MAS task, so they don't work very well on small datasets. In contrast, thanks to the superiority of our proposed BVLA and SDM, the key and novel information is easier to capture and align, therefore achieving much better performance on the smaller dataset.

To further elaborate on the advantage of our model in the case of small datasets, we conduct experiments on MM-AVS, which has the minimum amount of data among benchmark datasets. We select T5, BART, VG-T5, and VG-BART for comparison. As shown in Table 4, if we directly fine-tune the baselines on this dataset, severe over-fitting phenomena occurs in both the VG-T5 and VG-BART models. It can be further observed that the above two models can only coverage when first training on the larger How2 dataset, and then continuously fine-tuning on the smaller MM-AVS dataset. This can be attributed to the fact that it is hard for the baseline models to disentangle enough summary-worthy information from the general visual features if limited training data is provided. Instead, with the proposed summary-worthy mechanism, our model works surprisingly well on MM-AVS, no matter whether the model is first trained on the larger How2 dataset or not.

4.3 Ablation Study

Impacts of Different Components

We evaluate the effectiveness of different components by gradually removing three modules, i.e., the SDM, BVLA, and auxiliary visual encoder. *1) SWVR w/o D*: the self distillation is first removed; in such case, our model can only leverage the summary-related information from transcripts and becomes Fig.3(c). *2) SWVR w/o D+B*: The BVLA is further replaced with the uni-directional VLA, then our model can be summarized as Fig.3(b). Note that in this case, our model is not aware of any summary-worthy information. *3) SWVR w/o D+B+A*: Lastly, the introduced auxiliary visual encoder is dropped, and our model is degraded to the simplest form Fig.3(a).

We compare the three variants with the complete *SWVR* on the How2 dataset, and results are shown in Table 2. The general trend is that eliminating any of the modules (e.g., from SWVR w/o D to w/o D+B+A) will negatively impact the model's performance, confirming each component's benefits. Moreover, it is noticeable that the introduction of SMD can bring an improvement of 1.25% on average, suggesting the importance of equipping our visual encoder with the ability to capture summary-worthy values from the pseudo summary.

Method	**R-1**	**R-2**	**R-L**	**B-1**	**B-2**	**B-3**	**B-4**	**M**	**C**	**CF**
Transcript										
S2S†	58.6	40.6	53.8	55.2	45.6	39.9	35.8	27.6	2.35	-
PG†	57.2	39.5	52.8	55.3	45.6	39.8	35.7	26.8	2.13	-
TF†	59.0	41.0	54.3	56.6	46.7	40.8	36.6	27.7	2.30	-
T5*	62.8	45.0	57.5	60.5	50.4	44.2	39.6	30.6	2.76	61.7
BART*	64.0	46.4	58.9	62.4	52.6	46.4	42.0	31.7	2.97	63.9
Transcript + Video										
HA (RNN)†	60.3	42.5	55.7	57.2	47.7	41.8	37.5	28.8	2.48	-
HA (TF)†	60.2	43.1	55.9	58.6	48.3	43.3	38.1	28.9	2.51	-
MFFG (RNN)†	62.3	46.1	58.2	59.1	50.4	45.1	41.1	30.1	2.69	-
MFFG (TF)†	61.6	45.1	57.4	60.0	50.9	45.3	41.3	29.9	2.67	-
SFFG‡	63.2	46.4	58.9	61.5	52.3	46.5	42.4	31.6	2.74	-
VG-T5*	63.3	45.3	58.0	60.7	50.8	44.7	40.2	31.0	2.86	62.8
VG-BART*	68.0	51.4	63.3	65.2	56.3	50.4	46.0	34.0	3.28	69.7
Our Framework and the Variants										
SWVR	**69.1**	**53.1**	**64.4**	**68.9**	**60.0**	**54.3**	**50.0**	**36.8**	**3.58**	**72.8**
- w/o D	68.7	52.6	63.9	68.5	59.5	53.7	49.3	36.5	3.42	72.3
- w/o D+B	68.2	51.5	63.4	66.5	57.3	51.5	47.2	35.9	3.30	71.9
- w/o D+B+A	67.0	50.5	61.8	64.7	55.6	49.8	45.4	33.3	3.24	71.5

Table 2: Evaluation results of baselines and our proposed models on the How2 dataset, where R, B, M, C, and CF stand for ROUGE, BLEU, MENTOR, CIDEr, and Content F1, respectively. Results with †, ‡, and * marks are taken from [Liu *et al.*, 2020], [Liu *et al.*, 2022], and [Yu *et al.*, 2021a], respectively. Abbreviations D, B, and A stand for self distillation, bi-directional visual-language attention, and auxiliary visual encoder modules, respectively.

Method	**R-1**	**R-2**	**R-L**
Transcript			
S2S‡	46.01	25.16	39.98
T5	55.36	36.01	49.73
BART	56.56	37.22	50.44
Transcript + Video			
MFFG (RNN)‡	48.53	28.69	44.08
MFFG (TF)‡	49.27	28.26	43.41
SFFG‡	50.60	30.38	44.67
VG-T5	57.72	36.80	50.37
VG-BART	58.24	37.99	51.46
Ours	**59.92**	**40.75**	**53.81**

Table 3: Evaluation results of baselines and our proposed models on the How2-300 dataset. Results with ‡ mark are taken from [Liu *et al.*, 2022].

Method	**R-1**	**R-2**	**R-L**
Without Training on the How2 dataset			
T5	22.36	9.00	17.96
BART	22.97	9.02	17.65
VG-T5	12.96	1.12	10.82
VG-BART	13.95	1.20	11.59
Ours	23.67	10.23	18.97
First Training on the How2 dataset			
T5	23.46	9.95	18.
BART	23.36	9.56	17.97
VG-T5	25.11	10.50	19.80
VG-BART	25.04	11.00	20.06
Ours	**26.08**	**12.46**	**21.31**

Table 4: Evaluation results of baselines and our proposed models on the MM-AVS dataset.

Impacts of the Hyper-parameter of λ

In SDM, we introduce the hyper-parameter λ, which controls the trade-off between the cross-entropy loss and the self distillation loss. To analysis the sensitivity of λ, we manually select the values of λ from $\{0.01, 0.05, 0.1, 0.2, 0.5\}$. ROUGE-1 and BLEU-1 w.r.t λ on How2 datasets are illustrated in Figure 4. It can be observed that the performance of *SWVR* first increases and gets the peak when λ is not greater than 0.1. Afterward, the improvement is neutralized and even lost if λ becomes larger. This may be attributed to the fact that a simple linear layer in (11) may not mitigate the difference between different modalities, and thus, we speculate that a more well-

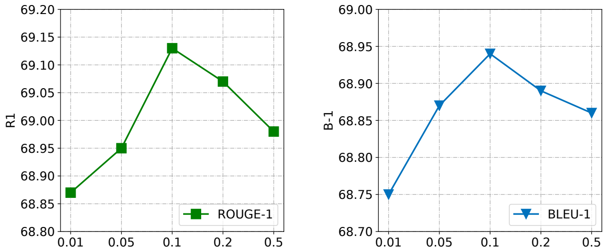
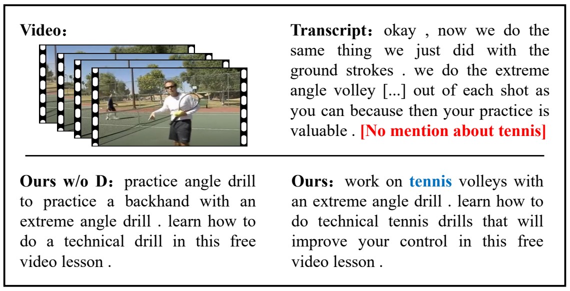


Figure 4: Impacts of λ on How2 dataset.



Video:

Transcript: okay , now we do the same thing we just did with the ground strokes . we do the extreme angle volley [...] out of each shot as you can because then your practice is valuable . **[No mention about tennis]**

Ours w/o D: practice angle drill to practice a backhand with an extreme angle drill . learn how to do a technical drill in this free video lesson .

Ours: work on **tennis** volleys with an extreme angle drill . learn how to do technical tennis drills that will improve your control in this free video lesson .

Figure 5: An example of case study. We show the generated summaries of model using summary-related and summary-worthy visual feature, respectively.

designed transformation may further discover the poentials of self distillation, which however, is beyond the scope of this paper.

4.4 Case Study

We conduct a case study to empirically exhibit the effectiveness of using summary-worthy visual information. To do so, we conduct experiments using two cases: one only exploits summary-worthy information from the transcript (see Fig.3(c)), while the other exploits summary-worthy information from both the transcript and the generated pseudo summary (Fig.3(d)). Data samples can be seen in Figure 5, the former case of only leveraging the summary-worthy values from the transcript fails to predict the ‘`tennis`’, which concept is only available in video frames. With the assistance of distilled knowledge from the pseudo summary, our model captures the novel concept, or we call it visual object (i.e., `tennis`), from the video frames and successfully generates a high-quality summary result.

4.5 Where to Adopt the BVLA and SDM

In this section, we aim to further investigate the optimal location where to insert the BVLA and SDM modules. As depicted in Table 5, in general, leveraging BVLA to inject the cross-modal features into the unimodal representation can significantly boost the model's performance. A similar phenomenon of the utilization of SDM can also be observed in Table 6. Furthermore, we observe that inserting the BVLA or SDM modules to the top layers can achieve the best performance. We speculate that the lower layers of the encoder may tend to capture the local and low-level semantics, which are usually modality-specific. This makes it challenging for the model to capture better cross-modal semantic interaction. In contrast, the top layers of the encoder may primarily exploit the global and high-level semantics, where the high-level abstract information is more accessible for the model to perform cross-modal interaction. This observation is also consistent with [Xu *et al.*, 2021]. Given this conclusion, our best model reported in Table 2 adopts the BVLA and SDM in {4,5,6}-th and {5,6}-th encoder layers, respectively.

Adoption of BVLA						**R-1**	**R-2**	**R-L**
1	2	3	4	5	6			
✗	✗	✗	✗	✗	✗	64.0	46.4	58.9
✓	✓	✓	✓	✓	✓	65.7	49.3	60.7
✗	✓	✓	✓	✓	✓	66.2	49.7	61.2
✗	✗	✓	✓	✓	✓	67.3	51.1	62.3
✗	✗	✗	✓	✓	✓	**68.7**	**52.6**	**63.9**
✗	✗	✗	✗	✓	✓	68.4	52.0	63.5
✗	✗	✗	✗	✗	✓	67.9	51.5	62.8

Table 5: How2 dataset performance of adopting BVLA at different locations in the encoder of *SWVR w/o D* (i.e., the self distillation mechanism is removed). ✓ indicates the adoption at a certain layer and ✗ indicates non-adoption.

Utilization of SDM						**R-1**	**R-2**	**R-L**
1	2	3	4	5	6			
✗	✗	✗	✗	✗	✗	65.7	49.3	60.7
✓	✓	✓	✓	✓	✓	66.1	49.7	60.9
✗	✓	✓	✓	✓	✓	66.5	50.0	61.4
✗	✗	✓	✓	✓	✓	67.0	50.4	61.7
✗	✗	✗	✓	✓	✓	67.3	50.9	62.1
✗	✗	✗	✗	✓	✓	**67.8**	**51.4**	**62.5**
✗	✗	✗	✗	✗	✓	67.5	50.8	61.9

Table 6: How2 dataset performance of utilizing SDM at different layers of the visual encoder. In this case, we force the model to adopt BVLA at all the locations. ✓ indicates the utilization of SDM a certain layer while ✗ indicates non-utilization.

5 Conclusion

In this paper, we propose to learn summary-worthy visual features to boost the MAS task. We introduce a bi-directional visual-language attention (BVLA) mechanism and a self distillation mechanism (SDM) to encourage the visual encoder to exploit the summary-worthy information from the textual data or the knowledge that distills from the pseudo summary. In addition, we enumerate almost all possible ways to combine and evaluate BVLA and SDM modules in the visual encoder. Experiments results show that our proposed method significantly outperforms all strong baselines on three public datasets. Further analysis demonstrates each component's effectiveness and suggests that the higher layers of the visual encoder are the optimal places to employ the BVLA and SMD.

Acknowledgments

We are grateful to ShuTing Cai for perfecting this article. This work is supported by the National Natural Science Foundation of China (No. 62276280, U1811264), Key R&D Program of Guangdong Province (No. 2018B010107005), Natural Science Foundation of Guangdong Province (No. 2021A1515012299), Science and Technology Program of Guangzhou (No. 202102021205).

References

[Ba *et al.*, 2016] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[Banko *et al.*, 2000] Michele Banko, Vibhu Mittal, and M. Witbrock. Headline generation based on statistical translation. In *Annual Meeting of the Association for Computational Linguistics*, 2000.

[Cherian *et al.*, 2022] Anoop Cherian, Chiori Hori, Tim K. Marks, and Jonathan Le Roux. (2.5+1)d spatio-temporal scene graphs for video question answering. In *AAAI Conference on Artificial Intelligence*, 2022.

[Clarke and Lapata, 2010] James Clarke and Mirella Lapata. Discourse constraints for document compression. *Computational Linguistics*, 36:411–441, 2010.

[Denkowski and Lavie, 2011] Michael J. Denkowski and Alon Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *WMT@EMNLP*, 2011.

[Erol *et al.*, 2003] Berna Erol, Dar-Shyang Lee, and Jonathan J. Hull. Multimodal summarization of meeting recordings. *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, 3:III–25, 2003.

[Evangelopoulos *et al.*, 2013] Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15:1553–1568, 2013.

[Fu *et al.*, 2021] Xiyan Fu, Jun Wang, and Zhenglu Yang. Mm-avs: A full-scale dataset for multi-modal summarization. In *North American Chapter of the Association for Computational Linguistics*, 2021.

[Hara *et al.*, 2018] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018.

[Jia *et al.*, 2021] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 2021.

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

[Kim *et al.*, 2021] Junyeong Kim, Sunjae Yoon, Dahyun Kim, and Chang D Yoo. Structured co-reference graph attention for video-grounded dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1789–1797, 2021.

[Kingma and Ba, 2014] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[Knight and Marcu, 2002] Kevin Knight and Daniel Marcu. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artif. Intell.*, 139:91–107, 2002.

[Lewis *et al.*, 2019] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Annual Meeting of the Association for Computational Linguistics*, 2019.

[Lewis *et al.*, 2020] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 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*, pages 7871–7880, 2020.

[Li *et al.*, 2017] Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. Multi-modal summarization for asynchronous collection of text, image, audio and video. In *Conference on Empirical Methods in Natural Language Processing*, 2017.

[Libovickỳ *et al.*, 2018] Jindrich Libovickỳ, Shruti Palaskar, Spandana Gella, and Florian Metze. Multimodal abstractive summarization for open-domain videos. *Visually Grounded Interaction and Language (ViGIL)*, pages 1–8, 2018.

[Lin and Hovy, 2003] Chin-Yew Lin and Eduard H. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *North American Chapter of the Association for Computational Linguistics*, 2003.

[Liu *et al.*, 2020] Nayu Liu, Xian Sun, Hongfeng Yu, Wenkai Zhang, and Guangluan Xu. Multistage fusion with forget gate for multimodal summarization in open-domain videos. In *Conference on Empirical Methods in Natural Language Processing*, 2020.

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

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

[Palaskar *et al.*, 2019] Shruti Palaskar, Jindřich Libovický, Spandana Gella, and Florian Metze. Multimodal abstractive summarization for how2 videos. In *Annual Meeting of the Association for Computational Linguistics*, 2019.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*, 2002.

[Paulus *et al.*, 2018] Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*, 2018.

[Qi *et al.*, 2020] Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. Prophetnet: Predicting future n-gram for sequence-to-sequencepre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2401–2410, 2020.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[Raffel *et al.*, 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

[Raghu *et al.*, 2021] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Neural Information Processing Systems*, 2021.

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

[See *et al.*, 2017] Abigail See, Peter J Liu, and Christopher D Manning. 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, 2017.

[Shah *et al.*, 2016] Rajiv Ratn Shah, Yi Yu, Akshay Verma, Suhua Tang, Anwar Dilawar Shaikh, and Roger Zimmermann. Leveraging multimodal information for event summarization and concept-level sentiment analysis. *Knowl. Based Syst.*, 108:102–109, 2016.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

[Tjondronegoro *et al.*, 2011] Dian Tjondronegoro, Xiaohui Tao, Johannes Sasongko, and Cher Han Lau. Multi-modal summarization of key events and top players in sports tournament videos. *2011 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 471–478, 2011.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Vedantam *et al.*, 2015] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015.

[Wang *et al.*, 2019] Kai Wang, Xiaojun Quan, and Rui Wang. Biset: Bi-directional selective encoding with template for abstractive summarization. In *Annual Meeting of the Association for Computational Linguistics*, 2019.

[Xu *et al.*, 2021] Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. Vlm: Task-agnostic video-language model pre-training for video understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4227–4239, 2021.

[Yu *et al.*, 2021a] Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 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, 2021.

[Yu *et al.*, 2021b] Tiezheng Yu, Zihan Liu, and Pascale Fung. Adaptsum: Towards low-resource domain adaptation for abstractive summarization. In *North American Chapter of the Association for Computational Linguistics*, 2021.

[Zhang *et al.*, 2019] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3712–3721, 2019.

[Zhang *et al.*, 2020] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR, 2020.

[Zhang *et al.*, 2022] Zhengkun Zhang, Xiaojun Meng, Yasheng Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. Unims: A unified framework for multimodal summarization with knowledge distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11757–11764, Jun. 2022.

[Zhou and Hovy, 2004] Liang Zhou and Eduard Hovy. Template-filtered headline summarization. In *Text Summarization Branches Out*, pages 56–60, Barcelona, Spain, July 2004. Association for Computational Linguistics.