SCCS: Semantics-Consistent Cross-domain Summarization via Optimal Transport Alignment

Jielin Qiu[1], **Jiacheng Zhu**[1], **Mengdi Xu**[1], **Franck Dernoncourt**[2],
Zhaowen Wang[2], **Trung Bui**[2], **Bo Li**[3], **Ding Zhao**[1], **Hailin Jin**[2]
[1]Carnegie Mellon University, [2]Adobe Research, [3]University of Illinois Urbana-Champaign
{jielinq,jzhu4,mengdixu,dingzhao}@andrew.cmu.edu, {dernonco,zhawang,bui,hljin}@adobe.com, lbo@illinois.edu

Abstract

Multimedia summarization with multimodal output (MSMO) is a recently explored application in language grounding. It plays an essential role in real-world applications, i.e., automatically generating cover images and titles for news articles or providing introductions to online videos. However, existing methods extract features from the whole video and article and use fusion methods to select the representative one, thus usually ignoring the critical structure and varying semantics with video/document. In this work, we propose a Semantics-Consistent Cross-domain Summarization (SCCS) model based on optimal transport alignment with visual and textual segmentation. Our method first decomposes both videos and articles into segments in order to capture the structural semantics, and then follows a cross-domain alignment objective with optimal transport distance, which leverages multimodal interaction to match and select the visual and textual summary. We evaluated our method on three MSMO datasets, and achieved performance improvement by 8% & 6% of textual and 6.6% &5.7% of video summarization, respectively, which demonstrated the effectiveness of our method in producing high-quality multimodal summaries.

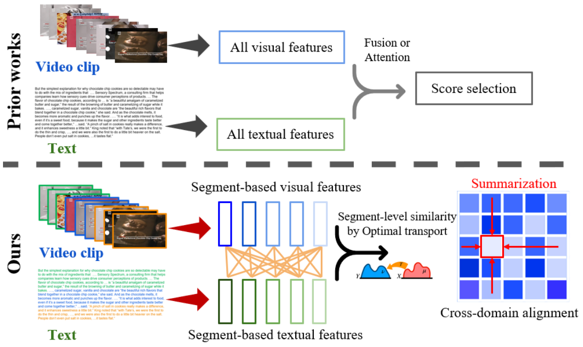


Figure 1: We proposed a segment-level cross-domain alignment model to preserve the structural semantics consistency within two domains for MSMO. We solve an optimal transport problem to optimize the cross-domain distance, which in turn finds the optimal match.

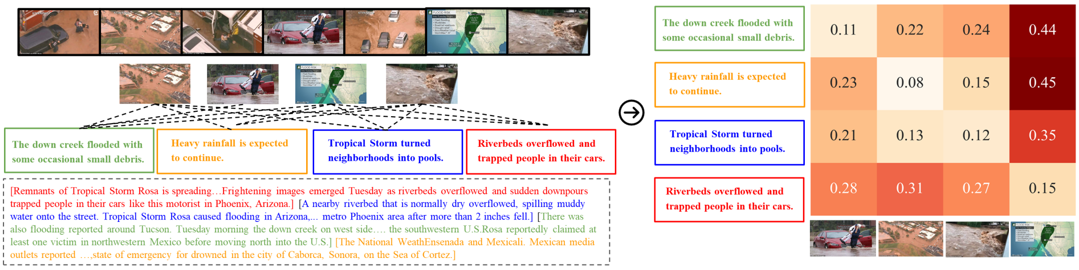
1 Introduction

New multimedia content in the form of short videos and corresponding text articles has become a significant trend in influential digital media. This popular media type has been shown to be successful in drawing users' attention and delivering essential information in an efficient manner. Multimedia summarization with multimodal output (MSMO) has recently drawn increasing attention. Different from traditional video or textual summarization (Gygli et al., 2014; Jadon and Jasim, 2020), where the generated summary is either a keyframe or textual description, MSMO aims at producing both visual and textual summaries simultaneously, making this task more complicated. Previous works addressed the MSMO task by processing the whole video and the whole article together which overlooked the structure and semantics of different domains (Duan et al., 2022; Haopeng et al., 2022; Sah et al., 2017; Zhu et al., 2018; Mingzhe et al., 2020; Fu et al., 2021, 2020).

The video and article can be regarded as being composed of several topics related to the main idea, while each topic specifically corresponds to one sub-idea. Thus, treating the whole video or article uniformly and learning a general representation ignores these structural semantics and easily leads to biased summarization. To address this problem, instead of learning averaged representations for the whole video & article, we focus on exploiting the original underlying structure. The comparison of our approach and previous works is illustrated in Figure 1. Our model first decomposes the video & article into segments to discover the content structure, then explores the cross-domain semantics relationship at the segment level. We believe this is a promising approach to exploit the *consistency* lie in the structural semantics between different domains.

Previous models applied attention or fusion

Findings of the Association for Computational Linguistics: ACL 2023, pages 1584–1601
July 9-14, 2023 ©2023 Association for Computational Linguistics



The down creek flooded with some occasional small debris.
Heavy rainfall is expected to continue.
Tropical Storm turned neighborhoods into pools.
Riverbeds overflowed and trapped people in their cars.

[Remnants of Tropical Storm Rosa is spreading…Frightening images emerged Tuesday as riverbeds overflowed and sudden downpours trapped people in their cars like this motorist in Phoenix, Arizona.] [A nearby riverbed that is normally dry overflowed, spilling muddy water onto the street. Tropical Storm Rosa caused flooding in Arizona,... metro Phoenix area after more than 2 inches fell.] [There was also flooding reported around Tucson. Tuesday morning the down creek on west side…. the southwestern U.S.Rosa reportedly claimed at least one victim in northwestern Mexico before moving north into the U.S.] [The National WeathEnsenada and Mexicali. Mexican media outlets reported …,state of emergency for drowned in the city of Caborca, Sonora, on the Sea of Cortez.]

The down creek flooded with some occasional small debris.	0.11	0.22	0.24	0.44
Heavy rainfall is expected to continue.	0.23	0.08	0.15	0.45
Tropical Storm turned neighborhoods into pools.	0.21	0.13	0.12	0.35
Riverbeds overflowed and trapped people in their cars.	0.28	0.31	0.27	0.15

Figure 2: A real example of the summarization process given by our SCCS method. Here we conduct OT-based cross-domain alignment to each keyframe-sentence pair, and a smaller OT distance means better alignment. (For example, the best-aligned text and image summary (0.08) delivers the flooding content clearly and comprehensively.)

mechanisms to compute image-text relevance scores, finding the best match of the sentences/images within the whole document/video, regardless of the context, which used one domain as an anchor. However, an outstanding anchor has more weight in selecting the corresponding pair. To overcome this, we believe the semantics structure is a crucial characteristic that can not be ignored. Based on this hypothesis, we propose Semantics-Consistent Cross-domain Summarization (SCCS), which explores segment-level cross-domain representations through Optimal Transport (OT) based multimodal alignment to generate both visual and textual summaries. We decompose the video/document into segments based on its semantic structure, then generate sub-summaries of each segment as candidates. We select the final summary from these candidates instead of a global search, so all candidates are in a fair competition arena.

Our contributions can be summarized as follow:

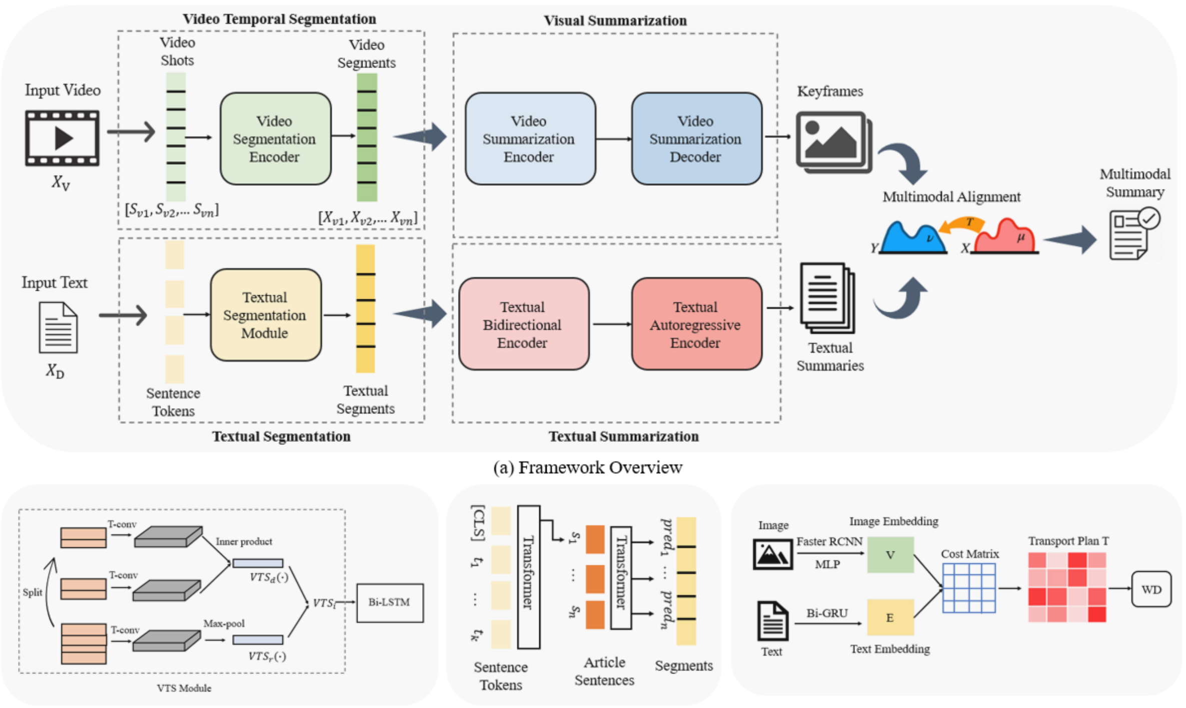
- We propose SCCS (Semantics-Consistent Cross-domain Summarization), a segment-level alignment model for MSMO tasks.
- Our method preserves the structural semantics and explores the cross-domain relationship through optimal transport to match and select the visual and textual summary.
- On three datasets, our method outperforms baselines in both textual and video summarization results qualitatively and quantitatively.
- Our method serves as a hierarchical MSMO framework and provides better interpretability via OT alignment. The OT coupling shows sparse patterns and specific temporal structure for the embedding vectors of ground-truth-matched video and text segments, providing interpretable learned representations.

Since MSMO generates both visual & textual summaries, We believe the optimal summary comes from the video and text pair that are both 1) semantically consistent, and 2) best matched globally in a cross-domain fashion. In addition, our framework is more computationally efficient as it conducts cross-domain alignment at the segment level instead of inputting whole videos/articles.

2 Related Work

Multimodal Alignment Aligning representations from different modalities is important in multimodal learning. Exploring the explicit relationship across vision and language has drawn significant attention (Wang et al., 2020a). Xu et al. (2015); Torabi et al. (2016); Yu et al. (2017) adopted attention mechanisms, Dong et al. (2021) composed pairwise joint representation, Chen et al. (2020b); Wray et al. (2019); Zhang et al. (2018) learned fine-grained or hierarchical alignment, Lee et al. (2018); Wu et al. (2019) decomposed the inputs into sub-tokens, Velickovic et al. (2018); Yao et al. (2018) adopted graph attention for reasoning, and Yang et al. (2021); Gutmann and Hyvärinen (2010); van den Oord et al. (2018); Radford et al. (2021) applied contrastive learning algorithms.

Multimodal Summarization Multimodal summarization explored multiple modalities, i.e., audio signals, video captions, transcripts, video titles, etc, for a summary generation. Otani et al. (2016); Yuan et al. (2019); Wei et al. (2018); Fu et al. (2020) learned the relevance or mapping in the latent space between different modalities. In addition to only generating visual summaries, Li et al. (2017); Atri et al. (2021); Zhu et al. (2018) generated textual summaries by taking audio, transcripts, or documents as input along with videos or images, using seq2seq model (Sutskever et al., 2014) or attention mechanism (Bahdanau et al., 2015). Recent trending on the MSMO task has also drawn much attention (Zhu et al., 2018; Mingzhe et al., 2020;



(a) Framework Overview

(b) Video Segmentation Encoder

(c) Textual Segmentation Module

(d) Multimodal Alignment Module

Figure 3: (a) The computational framework of the SCCS model, which takes multimodal inputs (videos & text documents) and generates multimodal summaries. The framework includes five modules: video temporal segmentation, visual summarization, textual segmentation, textual summarization, and multimodal alignment. (b) The structure of the video segmentation encoder. (c) The architecture of the textual segmentation module. (d) The multimodal alignment module for multimodal summaries.

Fu et al., 2021, 2020; Zhang et al., 2022). More related works are shown in Appendix B.

3 Methods

SCCS is a segment-level cross-domain semantics alignment model for the MSMO task, where MSMO aims at generating both visual and language summaries. We follow the problem setting in Mingzhe et al. (2020), for a multimedia source with documents and videos, the document $X_D = \{x_1, x_2, ..., x_d\}$ has d words, and the ground truth textual summary $Y_D = \{y_1, y_2, ..., y_g\}$ has g words. A corresponding video X_V is associated with the document in pair, and there exists a ground truth cover picture Y_V that can represent the most important information to describe the video. Our SCCS model generates both textual summaries Y'_D and video keyframes Y'_V.

SCCS consists of five modules, as shown in Figure 3(a): video temporal segmentation (Section 3.1), visual summarization (Section 3.3), textual segmentation (Section 3.2), textual summarization (Section 3.4), and cross-domain alignment (Section 3.5). Each module will be introduced in the following subsections.

3.1 Video Temporal Segmentation

Video temporal segmentation splits the original video into small segments, which summarization tasks build upon. The segmentation is formulated as a binary classification problem on the segment boundaries, similar to Rao et al. (2020). For a video X_V, the video segmentation encoder separates the video sequence into segments $[X_{v1}, X_{v2}, ..., X_{vm}]$, where n is the number of segments.

As shown in Figure 3(b), the video segmentation encoder contains a VTS module and a Bi-LSTM (Graves and Schmidhuber, 2005). Video X_V is first split into shots $[S_{v1}, S_{v2}, ..., S_{vn}]$ (Castellano, 2021), then the VTS module takes a clip of the video with $2\omega_b$ shots as input and outputs a boundary representation b_i. The boundary representation captures both differences and relations between the shots before and after. VTS consists of two branches, VTS$_d$ and VTS$_r$, as shown in Equation 1.

$$b_i = \mathrm{VTS}\left(\left[S_{vi-(\omega_b-1)},\cdots,S_{vi+\omega_b}\right]\right) = \begin{bmatrix} \mathrm{VTS}_d\left(\left[S_{vi-(\omega_b-1)},\cdots,\mathrm{P}_{vi}\right],\left[S_{v(i+1)},\cdots,S_{vi+\omega_b}\right]\right) \\ \mathrm{VTS}_r\left(\left[S_{vi-(\omega_b-1)},\cdots,P_{vi},S_{v(i+1)},\cdots,S_{vi+\omega_b}\right]\right) \end{bmatrix} \quad (1)$$

VTS$_d$ is modeled by two temporal convolution layers, each of which embeds the w_b shots before and after the boundary, respectively, following an inner product operation to calculate the differences. VTS$_r$ contains a temporal convolution layer followed by a max pooling, aiming at capturing the relations of the shots. It predicts a sequence binary labels $[p_{v1}, p_{v2}, ..., p_{vn}]$ based on the sequence of representatives $[b_1, b_2, ..., b_n]$. A Bi-LSTM (Graves and Schmidhuber, 2005) is used with stride $\omega_t/2$ shots to predict a sequence of coarse score $[s_1, s_2, ..., s_n]$, as shown in Equation 2,

$$[s_1, s_2, ..., s_n] = \text{Bi-LSTM}\left([b_1, b_2, \cdots, b_n]\right) \quad (2)$$

where $s_i \in [0, 1]$ is the probability of a shot boundary being a scene boundary. The coarse prediction $\hat{p}_{vi} \in \{0, 1\}$ indicates whether the i-th shot boundary is a scene boundary by binarizing s_i with a threshold τ, $\hat{p}_{vi} = \begin{cases} 1 & \text{if } s_i > \tau \\ 0 & \text{otherwise} \end{cases}$. The results with $\hat{p}_{vi} = 1$ result in the learned video segments $[X_{v1}, X_{v2}, ..., X_{vm}]$.

3.2 Textual Segmentation

The textual segmentation module takes the whole document or articles as input and splits the original input into segments based on context understanding. We used a hierarchical BERT as the textual segmentation module (Lukasik et al., 2020), which is the current state-of-the-art method. As shown in Figure 3(c), the textual segmentation module contains two-level transformer encoders, where the first-level encoder is for sentence-level encoding, and the second-level encoder is for article-level encoding. The hierarchical BERT starts by encoding each sentence with BERT$_{\text{LARGE}}$ independently, then the tensors produced for each sentence are fed into another transformer encoder to capture the representation of the sequence of sentences. All the sequences start with a [CLS] token to encode each sentence with BERT at the first level. If the segmentation decision is made at the sentence level, we use the [CLS] token as input for the second-level encoder. The [CLS] token representations from sentences are passed into the article encoder, which can relate the different sentences through cross-attention.

3.3 Visual Summarization

The visual summarization module generates visual keyframes from each video segment as its corresponding summary. We use an encoder-decoder architecture with attention as the visual summarization module (Ji et al., 2020), taking each video segment as input and outputting a sequence of keyframes. The encoder is a Bi-LSTM (Graves and Schmidhuber, 2005) to model the temporal relationship of video frames, where the input is $X = [x_1, x_2, ..., x_T]$ and the encoded representation is $E = [e_1, e_2, ...e_T]$. The decoder is a LSTM (Hochreiter and Schmidhuber, 1997) to generate output sequences $D = [d_1, d_2, ..., d_m]$. To exploit the temporal ordering across the entire video, an attention mechanism is used: $\text{E}_t = \sum_{i=1}^{m} \alpha_t^i e_i$, s.t. $\sum_{i=1}^{n} \alpha_t^i = 1$. Similar in Hochreiter and Schmidhuber (1997), the decoder function can be written as:

$$\begin{bmatrix} p\left(d_t \mid \{d_i \mid i < t\}, E_t\right) \\ s_t \end{bmatrix} = \psi\left(s_{t-1}, d_{t-1}, E_t\right) \quad (3)$$

where s_t is the hidden state, E_t is the attention vector at time t, α_t^i is the attention weight between the inputs and the encoder vector, ψ is the decoder function (LSTM). To obtain α_t^i, the relevance score γ_t^i is computed by $\gamma_t^i = \text{score}(s_{t-1}, e_i)$, where the score function decides the relationship between the i-th visual features e_i and the output scores at time t: $\gamma_t^i = e_i^T W_a s_{t-1}$, $\alpha_t^i = \exp(\gamma_t^i) / \sum_{j=1}^{m} \exp(\gamma_t^j)$.

3.4 Textual Summarization

Language summarization can produce a concise and fluent summary which should preserve the critical information and overall meaning. Our textual summarization module takes BART (Lewis et al., 2020) as the summarization model to generate abstractive textual summary candidates. BART is a denoising autoencoder that maps a corrupted document to the original document it was derived from. As in Figure 3(a), BART is an encoder-decoder Transformer pre-trained with a denoising objective on text. We take the fine-tuned BART on CNN and Daily Mail datasets for the summarization task (See et al., 2017b; Nallapati et al., 2016).

3.5 Cross-Domain Alignment via OT

Our cross-domain alignment (CDA) module learns the alignment between keyframes and textual summaries to generate the final multimodal summaries. Our alignment module is based on OT, which has been explored in several cross-domain tasks (Chen et al., 2020a; Yuan et al., 2020; Lu et al., 2021). More OT introductions can be found in Appendix A.

As shown in Figure 3(d), in CDA, the image features $V = \{\boldsymbol{v}_k\}_{k=1}^{K}$ are extracted from pre-trained

ResNet-101 (He et al., 2016) concatenated to faster R-CNN (Ren et al., 2015) as Yuan et al. (2020), where an image can be represented as a set of detected objects, each associated with a feature vector. For text features, every word is embedded as a feature vector and processed by a Bi-GRU (Cho et al., 2014) to account for context (Yuan et al., 2020). The extracted image and text embeddings are $\mathbf{V} = \{\boldsymbol{v}_i\}_1^K$, $\mathbf{E} = \{\boldsymbol{e}_i\}_1^M$, respectively.

As in Yuan et al. (2020), we take image and text sequence embeddings as two discrete distributions supported on the same feature representation space. Solving an OT transport plan between the two naturally constitutes a matching scheme to relate cross-domain entities (Yuan et al., 2020). To evaluate the OT distance, we compute a pairwise similarity between V and E using cosine distance:

$$C_{km} = C(e_k, v_m) = 1 - \frac{\boldsymbol{e}_k^T \boldsymbol{v}_k}{\|\boldsymbol{e}_k\|\,\|\boldsymbol{v}_m\|} \tag{4}$$

Then the OT can be formulated as:

$$\mathcal{L}_{\mathrm{OT}}(\mathbf{V}, \mathbf{E}) = \min_{\mathbf{T}} \sum_{k=1}^{K} \sum_{m=1}^{M} \mathbf{T}_{km} \mathbf{C}_{km} \tag{5}$$

where $\sum_m \mathbf{T}_{km} = \mu_k$, $\sum_k \mathbf{T}_{km} = v_m$, $\forall k \in [1, K]$, $m \in [1, M]$, $\mathbf{T} \in \mathbb{R}_+^{K \times M}$ is the transport matrix, d_k and d_m are the weight of $\boldsymbol{v}_k$ and $\boldsymbol{e}_m$ in a given image and text sequence, respectively. We assume the weight for different features to be uniform, i.e., $\mu_k = \frac{1}{K}, v_m = \frac{1}{M}$. The objective of optimal transport involves solving linear programming and may cause potential computational burdens since it has $O(n^3)$ efficiency. To solve this issue, we add an entropic regularization term equation (5), and the objective of our optimal transport distance becomes:

$$\mathcal{L}_{\mathrm{OT}}(\mathbf{V}, \mathbf{E}) = \min_{\mathbf{T}} \sum_{k=1}^{K} \sum_{m=1}^{M} \mathbf{T}_{km} \mathbf{C}_{km} + \lambda H(\mathbf{T}) \tag{6}$$

where $H(\mathbf{T}) = \sum_{i,j} \mathbf{T}_{i,j} \log \mathbf{T}_{i,j}$ is the entropy, and λ is the hyperparameter that balance the effect of the entropy term. Thus, we are able to apply the celebrated Sinkhorn algorithm (Cuturi, 2013) to efficiently solve the above equation in $O(nlogn)$. The optimal transport distance computed via the Sinkhorn algorithm is differentiable and can be implemented by Flamary et al. (2021). The algorithm is shown in Algorithm 1, where β is a hyperparameter, $\mathbf{C}$ is the cost matrix, $\odot$ is Hadamard product, $< \cdot, \cdot >$ is Frobenius dot-product, matrices are in bold, the rest are scalars.

Algorithm 1 Compute Alignment Distance

```
Input: $\mathbf{E} = \{\boldsymbol{e}_i\}_1^M, \mathbf{V} = \{\boldsymbol{v}_i\}_1^K, \beta$
$\mathbf{C} = C(\mathbf{V}, \mathbf{E}), \sigma \leftarrow \frac{1}{m}\mathbf{1_m}, \mathbf{T}^{(1)} \leftarrow \mathbf{1}\mathbf{1}^T$
$\mathbf{G}_{ij} \leftarrow \exp\left(-\frac{\boldsymbol{C}_{ij}}{\beta}\right)$
for t = 1,2,3,...,N do
    $\boldsymbol{Q} \leftarrow \boldsymbol{G} \odot \mathbf{T}^{(t)}$
    $\boldsymbol{\delta} \leftarrow \frac{1}{K\boldsymbol{Q}\boldsymbol{\sigma}}, \boldsymbol{\sigma} \leftarrow \frac{1}{M\boldsymbol{Q}^T\boldsymbol{\delta}}$
    $\mathbf{T}^{(t+1)} \leftarrow \mathrm{diag}(\boldsymbol{\delta})\boldsymbol{Q}\,\mathrm{diag}(\boldsymbol{\sigma})$
end for
Dis $=< C^T, T >$
```

3.6 Multimodal Summaries

During training the alignment module, the Wasserstein distance (WD) between each keyframe-sentence pair of all the visual & textual summary candidates is computed, where the best match is selected as the final multimodal summaries.

4 Datasets and Baselines

4.1 Datasets

We evaluated our models on three datasets: VMSMO dataset, Daily Mail dataset, and CNN dataset from Mingzhe et al. (2020); Fu et al. (2021, 2020). The VMSMO dataset contains 184,920 samples, including articles and corresponding videos. Each sample is assigned with a textual summary and a video with a cover picture. We adopted the available data samples from Mingzhe et al. (2020). The Daily Mail dataset contains 1,970 samples, and the CNN dataset contains 203 samples, which include video titles, images, and captions, similar to Hermann et al. (2015). For data splitting, we take the same experimental setup as Mingzhe et al. (2020) for the VMSMO dataset. For the Daily Mail dataset and CNN dataset, we split the data by 70%, 10%, and 20% for train, validation, and test sets, respectively, same as Fu et al. (2021, 2020).

4.2 Baselines

We select state-of-the-art MSMO baselines and representative pure video & textual summarization baselines for comparison. For the VMSMO dataset, we compare our method with (i) multimodal summarization baselines (MSMO, MOF (Zhu et al., 2018, 2020), and DIMS (Mingzhe et al., 2020), (ii) video summarization baselines (Synergistic (Guo et al., 2019) and PSAC (Li et al., 2019)), and (iii) textual summarization baselines (Lead (See et al., 2017a), TextRank (Mihalcea and Tarau, 2004), PG (See et al., 2017b), Unified (Hsu et al., 2018), and GPG (Shen et al., 2019)). For Daily Mail and

CNN datasets, we compare our method with (i) multimodal baselines (MSMO (Zhu et al., 2018), Img+Trans (Hori et al., 2019), TFN (Zadeh et al., 2017), HNNattTI (Chen and Zhuge, 2018), and M^{2}SM (Fu et al., 2021, 2020)), (ii) video summarization baselines (VSUMM (De Avila et al., 2011) and DR-DSN (Zhou et al., 2018a)), and (iii) textual summarization baselines (Lead3 (See et al., 2017a), NN-SE (Cheng and Lapata, 2016), BART (Lewis et al., 2020), T5 (Raffel et al., 2019), and Pegasus (Zhang et al., 2019a)). More details about the baselines are introduced in Appendix C.

5 Experiments

5.1 Experimental Setting and Implementation

For the VTS module, we used the same model setting as Rao et al. (2020); Castellano (2021) and the same data splitting setting as Mingzhe et al. (2020); Fu et al. (2021, 2020) in the training process.

The visual summarization model is pre-trained on the TVSum (Song et al., 2015) and SumMe (Gygli et al., 2014) datasets. TVSum dataset contains 50 edited videos downloaded from YouTube in 10 categories, and SumMe dataset consists of 25 raw videos recording various events. Frame-level importance scores for each video are provided for both datasets and used as ground-truth labels. The input visual features are extracted from pre-trained GoogLeNet on ImageNet, where the output of the pool5 layer is used as visual features.

For the textual segmentation module, due to the quadratic computational cost of transformers, we reduce the BERT's inputs to 64-word pieces per sentence and 128 sentences per document as Lukasik et al. (2020). We use 12 layers for both the sentence and the article encoders, for a total of 24 layers. In order to use the BERT$_{\text{BASE}}$ checkpoint, we use 12 attention heads and 768-dimensional word-piece embeddings. The hierarchical BERT model is pre-trained on the Wiki-727K dataset (Koshorek et al., 2018), which contains 727 thousand articles from a snapshot of the English Wikipedia. We used the same data splitting method as Koshorek et al. (2018).

For textual summarization, we adopted the pre-trained BART model from Lewis et al. (2020), which contains 1024 hidden layers and 406M parameters and has been fine-tuned using CNN and Daily Mail datasets.

In the cross-domain alignment module, the feature extraction and alignment module is pretrained by MS COCO dataset (Lin et al., 2014) on the image-text matching task. We added the OT loss as a regularization term to the original matching loss to align the image and text more explicitly.

5.2 Evaluation Metrics

The quality of generated textual summary is evaluated by standard Rouge F1 (Lin, 2004) following previous works (See et al., 2017b; Chen et al., 2018; Mingzhe et al., 2020). ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L) refer to the overlap of unigram, bigrams, and the longest common subsequence between the decoded summary and the reference, respectively (Lin, 2004). Due to the limitation of ROUGE, we also adopt BertScore (Zhang et al., 2019b) for evaluation.

For the VMSMO dataset, the quality of the chosen cover frame is evaluated by mean average precision (MAP) and recall at position $(R_n@k)$ (Zhou et al., 2018c; Tao et al., 2019), where $(R_n@k)$ measures if the positive sample is ranked in the top k positions of n candidates. For the Daily Mail dataset and CNN dataset, we calculate the cosine image similarity (Cos) between image references and the extracted frames (Fu et al., 2021, 2020).

5.3 Results and Discussion

The comparison results on the VMSMO dataset of multimodal, video, and textual summarization are shown in Table 1. Synergistic and PSAC are pure video summarization approaches, which did not perform as well as multimodal methods, like MOF or DIMS, which means taking additional modality into consideration actually helps to improve the quality of the generated video summaries. Table 1 also shows the absolute performance improvement or decrease compared with the MSMO baseline, where the improvements are marked in red and decreases in blue. Overall, our method shows the highest absolute performance improvement than the previous methods on both textual and video summarization results. Our method shows the ability to preserve the structural semantics and is able to learn the alignment between keyframes and textual deceptions, which shows better performance than the previous ones. If comparing the quality of generated textual summaries, our method still outperforms the other multimodal baselines, like MSMO, MOF, DIMS, and also traditional textual summarization methods, like Lead, TextRank, PG, Unified, and GPG, showing the alignment obtained

Table 1: Comparison with multimodal baselines on the VMSMO dataset. The absolute performance comparison with the baseline MSMO method is marked in red (better) and blue (worse).

Category	Methods	Textual			Video			
		R-1	R-2	R-L	MAP	R_{10}@1	R_{10}@2	R_{10}@5
Video	Synergistic	–	–	–	0.558	0.444	0.557	0.759
	PSAC	–	–	–	0.524	0.363	0.481	0.730
Textual	Lead	16.2	5.3	13.9	–	–	–	–
	TextRank	13.7	4.0	12.5	–	–	–	–
	PG	19.4	6.8	17.4	–	–	–	–
	Unified	23.0	6.0	20.9	–	–	–	–
	GPG	20.1	4.5	17.3	–	–	–	–
Multimodal	MSMO	20.1	4.6	17.3	0.554	0.361	0.551	0.820
	MOF	21.3 (↑ 0.8)	5.7 (↑ 1.1)	17.9 (↑ 0.6)	0.615 (↑ 0.061)	0.455 (↑ 0.094)	0.615 (↑ 0.064)	0.817 (↓ -0.003)
	DIMS	25.1 (↑ 5.0)	9.6 (↑ 5.0)	23.2 (↑ 5.9)	0.654 (↑ 0.100)	0.524 (↑ 0.163)	0.634 (↑ 0.083)	0.824 (↑ 0.004)
Ours	Ours-textual	26.2	9.6	24.1	–	–	–	–
	Ours-video	–	–	–	0.678	0.561	0.642	0.863
	Ours	**27.1** (↑ 7.0)	**9.8** (↑ 5.2)	**25.4** (↑ 8.1)	**0.697** (↑ 0.143)	**0.582** (↑ 0.221)	**0.688** (↑ 0.137)	**0.895** (↑ 0.075)

by optimal transport can help to identify the cross-domain inter-relationships.

In Table 2, we show the comparison results with multimodal baselines on the Daily Mail and CNN datasets. We can see that for the CNN datasets, our method shows competitive results with Img+Trans, TFN, HNNattTI, and M^2SM on the quality of generated textual summaries. While on the Daily Mail dataset, our approach showed better performance on both textual summaries and visual summaries. We also compare with the traditional pure video summarization baselines and pure textual summarization baselines on the Daily Mail dataset, and the results are shown in Table 2. We can find that our approach performed competitive results compared with NN-SE and M^2SM for the quality of the generated textual summary. For visual summarization comparison, we can find that the quality of generated visual summary by our approach still outperforms the other visual summarization baselines. Still, we also provide absolute performance comparison with baseline MSMO (Zhu et al., 2018), as shown in Table 2, our model achieved the highest performance improvement in both Daily Mail and CNN datasets compared with previous baselines. If comparing the quality of generated textual summaries with language model (LM) baselines, our method also outperforms T5, Pegasus, and BART.

5.4 Human Evaluation

To provide human evaluation results, we asked 5 people (recruited from the institute) to score the results generated by different approaches of CNN and DailyMail datasets. We asked the human judges to score the results of 5 models: MSMO, TFN, HNNattTI, M^{2}SM, and SCCS, as 1-5, where 5 represents the best results. We averaged the voting results from 5 human judges. The performances of 5 models are listed in Table 3, showing the result by SCCS is better than the baselines.

Table 3: Human evaluation results.

Method	MSMO	TFN	HNNattTI	M^{2}SM	SCCS
Score	1.84	2.36	3.24	3.4	**4.16**

5.5 Factual Consistency Evaluation

Factual consistency is used as another important evaluation criterion for evaluating summarization results (Honovich et al., 2022). For factual consistency, we adopted the method in Xie et al. (2021) and followed the same setting. The same human annotators from Sec 5.4 provided human judgments. We report Pearson correlation coefficient Coe$_P$ here. The results of MSMO, Img+Trans, TFN, HNNaatTI, M^{2}SM, and ours, are shown in Table 4. In summary, our methods show better results than baselines on factual consistency evaluations.

Table 4: Factual consistency evaluation results.

Datasets	MSMO	Img+Trans	TFN	HNNattTI	M^{2}SM	SCCS
CNN	40.12	41.23	41.52	42.33	42.59	**44.37**
DailyMail	50.31	50.65	50.72	51.37	51.69	**53.16**

5.6 Ablation Study

To evaluate each component's performance, we performed ablation experiments on different modalities and different datasets. For the VMSMO dataset, we compare the performance of using only visual information, only textual information, and multimodal information. The comparison result is shown in Table 1. We also carried out experiments on different modalities using Daily Mail dataset to show the performance of unimodal and multimodal components, and the results are shown in Table 2.

For ablation results, when only textual data is available, we adopt BERT (Devlin et al., 2019) to generate text embeddings and K-Means clustering

Table 2: Comparisons of multimodal baselines on the Daily Mail and CNN datasets. The absolute performance comparison with the baseline MSMO method is marked in red (better) and blue (worse).

Category	Methods	CNN dataset				Daily Mail dataset				
		R-1	R-2	R-L	BertScore	R-1	R-2	R-L	BertScore	Cos(%)
Video	VSUMM	–	–	–	–	–	–	–	–	68.74
	DR-DSN	–	–	–	–	–	–	–	–	68.69
Textual	Lead3	–	–	–	–	41.07	17.87	30.90	–	–
	NN-SE	–	–	–	–	41.22	18.15	31.22	–	–
	T5	27.31	8.78	18.22	–	42.32	18.23	33.45	–	–
	Pegasus	27.28	8.83	18.59	–	43.01	18.63	33.54	–	–
	BART	27.50	8.76	18.83	–	42.49	18.67	33.92	–	–
Multimodal	MSMO	26.83	8.11	18.34	12.13	35.38	14.79	25.41	16.25	69.17
	Img+Trans	27.04 (↑ 0.21)	8.29 (↑ 0.18)	18.54 (↑ 0.20)	12.35 (↑ 0.22)	39.28 (↑ 3.90)	16.64 (↑ 1.85)	28.53 (↑ 3.12)	16.43 (↑ 0.18)	-
	TFN	27.68 (↑ 0.85)	8.69 (↑ 0.58)	18.71 (↑ 0.37)	12.59 (↑ 0.46)	39.37 (↑ 3.99)	16.38 (↑ 1.59)	28.09 (↑ 2.68)	16.71 (↑ 0.46)	-
	HNNattTI	27.61 (↑ 0.78)	8.74 (↑ 0.63)	18.64 (↑ 0.30)	12.67 (↑ 0.54)	39.58 (↑ 4.20)	16.71 (↑ 1.92)	29.04 (↑ 3.63)	16.79 (↑ 0.54)	68.76 (↓ -0.41)
	M^{2}SM	27.81 (↑ 0.98)	8.87 (↑ 0.76)	18.73 (↑ 0.39)	12.72 (↑ 0.59)	41.73 (↑ 6.35)	18.59 (↑ 3.80)	31.68 (↑ 6.27)	16.93 (↑ 0.68)	69.22 (↑ 0.05)
Ours	Ours-textual	–	–	–	12.68	40.28	17.93	31.89	16.98	–
	Ours-video	–	–	–	–	–	–	–	–	70.56
	Ours-Multimodal	**28.02** (↑ 1.19)	**8.94** (↑ 0.83)	**18.89** (↑ 0.55)	**13.21** (↑ 1.08)	**44.52** (↑ 9.14)	**19.87** (↑ 5.08)	**35.79** (↑ 10.38)	**17.45** (↑ 1.20)	**73.19** (↑ 4.02)

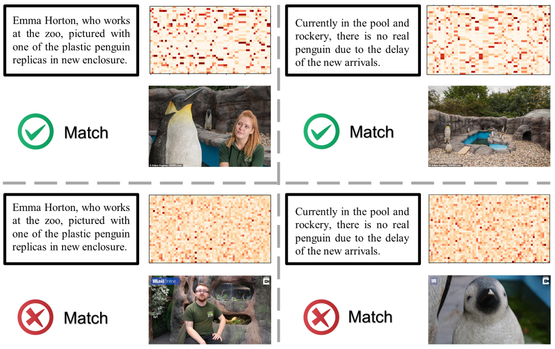
to identify sentences closest to the centroid for textual summary selection. While if only video data is available, we solve the visual summarization task in an unsupervised manner, using K-Means clustering to cluster frames using the image histogram and then select the best frame from clusters based on the variance of laplacian as the visual summary.

From Table 1 and Table 2, we can find that multimodal methods outperform unimodal approaches, showing the effectiveness of exploring the relationship and taking advantage of the cross-domain alignments of generating high-quality summaries.

5.7 Interpretation

To show a deeper understanding of the multimodal alignment between the visual domain and language domain, we compute and visualize the transport plan to provide an interpretation of the latent representations, which is shown in Figure 4. When we are regarding the extracted embedding from both text and image spaces as the distribution over their corresponding spaces, we expect the optimal transport coupling to reveal the underlying similarity and structure. Also, the coupling seeks sparsity, which further helps to explain the correspondence between the text and image data.

Figure 4 shows comparison results of matched image-text pairs and non-matched ones. The top two pairs are shown as matched pairs, where there is an overlap between the image and the corresponding sentence. The bottom two pairs are shown as non-matched ones, where the overlapping of meaning between the image and text is relatively small. The correlation between the image domain and the language domain can be easily interpreted by the learned transport plan matrix. In specific, the optimal transport coupling shows the pattern of sequentially structured knowledge. However, for non-matched image-sentences pairs, the estimated couplings are relatively dense and barely contain any informative structure. As shown in Figure 4, we can find that the transport plan learned in the cross-domain alignment module demonstrates a way to align the features from different modalities to represent the key components. The visualization of the transport plan contributes to the interpretability of the proposed model, which brings a clear understanding of the alignment module.



Emma Horton, who works at the zoo, pictured with one of the plastic penguin replicas in new enclosure.

Currently in the pool and rockery, there is no real penguin due to the delay of the new arrivals.

✓ Match
✓ Match
Emma Horton, who works at the zoo, pictured with one of the plastic penguin replicas in new enclosure.
Currently in the pool and rockery, there is no real penguin due to the delay of the new arrivals.
⊗ Match
⊗ Match

Figure 4: The OT coupling shows sparse patterns and specific temporal structure for the embedding vectors of ground-truth-matched video and text segments.

6 Conclusion

In this work, we proposed SCCS, a segment-level Semantics-Consistent Cross-domain Summarization model for the MSMO task. Our model decomposed the video & article into segments based on the content to preserve the structural semantics, and explored the cross-domain semantics relationship via optimal transport alignment at the segment level. The experimental results on three MSMO datasets show that SCCS outperforms previous summarization methods. We further provide interpretation by the OT coupling. Our approach provides a new direction for the MSMO task, which can be extended to many real-world applications.

7 Limitations

Due to the absence of large evaluation databases, we only evaluated our method on three publicly available datasets that can be used for the MSMO task. The popular video databases, i.e., COIN and Howto100M datasets, can not be used in our task, since they lack narrations and key-step annotation. So a large evaluation database is highly needed for evaluating the performance of MSMO approaches.

As the nature of the summarization task, human preference has an inevitable influence on the performance, since the ground-truth labels were provided by human annotators. It's somehow difficult to quantitatively specify the quality of the summarization result, and current widely used evaluation metrics may not reflect the performance of the results very well. So we are seeking some new directions to find another idea for quality evaluation.

The current setting is short videos & short documents, due to the constrain of available data. To extend the current MSMO to a more general setting, i.e., much longer videos or documents, new datasets should be collected. However, this requires huge human effort in annotating and organizing a high-value dataset, which is extremely time-consuming and labor-intensive. Nevertheless, we believe the MSMO task is promising and can provide valuable solutions to many real-world problems. So if such a dataset is collected, we believe it could significantly boost the research in this field.

8 Ethics Statement

Our work aims at providing a better user experience when exploring online multimedia, and there is no new dataset collected. To the best of our knowledge, this application does not involve ethical issues, and we do not foresee any harmful uses of this study.

9 Acknowledgements

The research is partially supported by Adobe Research and the DARPA ADAPTER program. We also sincerely appreciate the suggestions and feedback from Daniel Fried.

References

Sathyanarayanan N. Aakur and Sudeep Sarkar. 2019. A perceptual prediction framework for self supervised event segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1197–1206.

Sawsan Alqahtani, Garima Lalwani, Yi Zhang, Salvatore Romeo, and Saab Mansour. 2021. Using optimal transport as alignment objective for fine-tuning multilingual contextualized embeddings. In *EMNLP*.

Evlampios E. Apostolidis, E. Adamantidou, Alexandros I. Metsai, Vasileios Mezaris, and I. Patras. 2021. Video summarization using deep neural networks: A survey. *Proceedings of the IEEE*, 109:1838–1863.

Yash Kumar Atri, Shraman Pramanick, Vikram Goyal, and Tanmoy Chakraborty. 2021. See, hear, read: Leveraging multimodality with guided attention for abstractive text summarization. *ArXiv*, abs/2105.09601.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

David M. Blei, A. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Brandon Castellano. 2021. Intelligent scene cut detection and video splitting tool. `https://bcastell.com/projects/PySceneDetect/`.

Harr Chen, S. R. K. Branavan, Regina Barzilay, and David R. Karger. 2009. Global models of document structure using latent permutations. In *NAACL*.

Jingqiang Chen and Hai Zhuge. 2018. Abstractive text-image summarization using multi-modal attentional hierarchical rnn. In *EMNLP*, pages 4046–4056.

Liqun Chen, Zhe Gan, Y. Cheng, Linjie Li, L. Carin, and Jing jing Liu. 2020a. Graph optimal transport for cross-domain alignment. *ICML*.

Liqun Chen, Yizhe Zhang, Ruiyi Zhang, Chenyang Tao, Zhe Gan, Haichao Zhang, Bai Li, Dinghan Shen, Changyou Chen, and Lawrence Carin. 2019. Improving sequence-to-sequence learning via optimal transport. *ArXiv*, abs/1901.06283.

Shixing Chen, Xiaohan Nie, David D. Fan, Dongqing Zhang, Vimal Bhat, and Raffay Hamid. 2021. Shot contrastive self-supervised learning for scene boundary detection. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9791–9800.

Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. 2020b. Fine-grained video-text retrieval with hierarchical graph reasoning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10635–10644.

Xiuying Chen, Shen Gao, Chongyang Tao, Yan Song, Dongyan Zhao, and Rui Yan. 2018. Iterative document representation learning towards summarization with polishing. In *EMNLP*.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *ACL*, pages 484–494.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *EMNLP*.

Freddy Y. Y. Choi. 2000. Advances in domain independent linear text segmentation. In *ANLP*.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300.

Sandra Eliza Fontes De Avila, Ana Paula Brandão Lopes, Antonio da Luz Jr, and Arnaldo de Albuquerque Araújo. 2011. Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, 32(1):56–68.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. 2021. Dual encoding for video retrieval by text. *IEEE transactions on pattern analysis and machine intelligence*, PP.

Jiali Duan, Liqun Chen, Son Thai Tran, Jinyu Yang, Yi Xu, Belinda Zeng, and Trishul M. Chilimbi. 2022. Multi-modal alignment using representation codebook. *ArXiv*, abs/2203.00048.

Rémi Flamary et al. 2021. Pot: Python optimal transport.

Xiyan Fu, Jun Wang, and Zhenglu Yang. 2020. Multimodal summarization for video-containing documents. *ArXiv*, abs/2009.08018.

Xiyan Fu, Jun Wang, and Zhenglu Yang. 2021. Mm-avs: A full-scale dataset for multi-modal summarization. In *NAACL*.

Goran Glavas, Federico Nanni, and Simone Paolo Ponzetto. 2016. Unsupervised text segmentation using semantic relatedness graphs. In **SEMEVAL*.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18 5-6:602–10.

Dalu Guo, Chang Xu, and Dacheng Tao. 2019. Image-question-answer synergistic network for visual dialog. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10426–10435.

Michael U Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*.

Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. 2014. Creating summaries from user videos. In *ECCV*.

Li Haopeng, Ke Qiuhong, Gong Mingming, and Zhang Rui. 2022. Video summarization based on video-text modelling.

Ahmed Hassanien, Mohamed A. Elgharib, Ahmed A. S. Seleim, Mohamed Hefeeda, and Wojciech Matusik. 2017. Large-scale, fast and accurate shot boundary detection through spatio-temporal convolutional neural networks. *ArXiv*, abs/1705.03281.

Eman Hato and Matheel Emaduldeen Abdulmunem. 2019. Fast algorithm for video shot boundary detection using surf features. *2019 2nd Scientific Conference of Computer Sciences (SCCS)*, pages 81–86.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9:1735–1780.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Y. Matias. 2022. True: Re-evaluating factual consistency evaluation. In *Workshop on Document-grounded Dialogue and Conversational Question Answering*.

Chiori Hori et al. 2019. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *ICASSP*, pages 2352–2356.

Wan Ting Hsu, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. *ArXiv*, abs/1805.06266.

Shruti Jadon and Mahmood Jasim. 2020. Unsupervised video summarization framework using keyframe extraction and video skimming. In *ICCCA*, pages 140–145.

Zhong Ji, Kailin Xiong, Yanwei Pang, and Xuelong Li. 2020. Video summarization with attention-based encoder–decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 30:1709–1717.

Johannes Klicpera, Marten Lienen, and Stephan Günnemann. 2021. Scalable optimal transport in high dimensions for graph distances, embedding alignment, and more. *ArXiv*, abs/2107.06876.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. In *NAACL*.

Hilde Kuehne, Alexander Richard, and Juergen Gall. 2020. A hybrid rnn-hmm approach for weakly supervised temporal action segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:765–779.

Colin S. Lea, Michael D. Flynn, René Vidal, Austin Reiter, and Gregory Hager. 2017. Temporal convolutional networks for action segmentation and detection. In *CVPR*.

John Lee, Max Dabagia, Eva L. Dyer, and Christopher J. Rozell. 2019. Hierarchical optimal transport for multimodal distribution alignment. *ArXiv*, abs/1906.11768.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. *ArXiv*, abs/1803.08024.

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 *ACL*.

Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2017. Multi-modal summarization for asynchronous collection of text, image, audio and video. In *EMNLP*.

J. Li, Aixin Sun, and Shafiq R. Joty. 2018. Segbot: A generic neural text segmentation model with pointer network. In *IJCAI*.

Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. 2019. Beyond rnns: Positional self-attention with co-attention for video question answering. In *AAAI*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3730–3740, Hong Kong, China. Association for Computational Linguistics.

W. Lu, Yiqiang Chen, Jindong Wang, and Xin Qin. 2021. Cross-domain activity recognition via substructural optimal transport. *Neurocomputing*.

Michal Lukasik, Boris Dadachev, Kishore Papineni, and Gonçalo Simões. 2020. Text segmentation by cross segment attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4707–4716, Online. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *EMNLP*.

Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *NAACL*.

Li Mingzhe, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2020. Vmsmo: Learning to generate multimodal summary for video-based news articles. In *EMNLP*.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *AAAI*.

Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, Çaglar Gülçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *CoNLL*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. In *NAACL*, pages 1747–1759.

Ana Sofia Nicholls. 2021. A neural model for text segmentation.

Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. 2016. Video summarization using deep semantic features. *ArXiv*, abs/1609.08758.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Yair Poleg, Chetan Arora, and Shmuel Peleg. 2014. Temporal segmentation of egocentric videos. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2537–2544.

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. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. 2020. A local-to-global approach to multi-modal movie scene segmentation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10143–10152.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149.

Martin Riedl and Chris Biemann. 2012. Topictiling: A text segmentation algorithm based on lda. In *ACL 2012*.

Shagan Sah, Sourabh Kulhare, Allison Gray, Subhashini Venugopalan, Emily Tucker Prud'hommeaux, and Raymond W. Ptucha. 2017. Semantic text summarization of long videos. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 989–997.

M. Saquib Sarfraz et al. 2021. Temporally-weighted hierarchical clustering for unsupervised action segmentation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11220–11229.

A. See, Peter J. Liu, and Christopher D. Manning. 2017a. Get to the point: Summarization with pointer-generator networks. *ArXiv*, abs/1704.04368.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017b. Get to the point: Summarization with pointergenerator networks. In *ACL*.

Xiaoyu Shen, Yang Zhao, Hui Su, and Dietrich Klakow. 2019. Improving latent alignment in text summarization by generalizing the pointer generator. In *EMNLP*.

Panagiotis Sidiropoulos, Vasileios Mezaris, Yiannis Kompatsiaris, Hugo Meinedo, Miguel M. F. Bugalho, and Isabel Trancoso. 2011. Temporal video segmentation to scenes using high-level audiovisual features. *IEEE Transactions on Circuits and Systems for Video Technology*, 21:1163–1177.

Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Ndedi Monekosso, and Paolo Remagnino. 2018. Superframes, a temporal video segmentation. *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 566–571.

Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. Tvsum: Summarizing web videos using titles. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5179–5187.

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

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. Abstractive document summarization with a graph-based attentional neural model. In *ACL*, pages 1171–1181.

Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019. Multi-representation fusion network for multi-turn response selection in retrieval-based chatbots. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*.

Atousa Torabi, Niket Tandon, and Leonid Sigal. 2016. Learning language-visual embedding for movie understanding with natural-language. *ArXiv*, abs/1609.08124.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio', and Yoshua Bengio. 2018. Graph attention networks. *ArXiv*, abs/1710.10903.

Cédric Villani. 2003. Topics in optimal transportation.

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2019. Temporal segment networks for action recognition in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:2740–2755.

Qinxin Wang, Haochen Tan, Sheng Shen, Michael W. Mahoney, and Zhewei Yao. 2020a. An effective framework for weakly-supervised phrase grounding. *ArXiv*, abs/2010.05379.

Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward fast and accurate neural discourse segmentation. In *EMNLP*.

Zhenzhi Wang, Ziteng Gao, Limin Wang, Zhifeng Li, and Gangshan Wu. 2020b. Boundary-aware cascade networks for temporal action segmentation. In *ECCV*.

Huawei Wei, Bingbing Ni, Yichao Yan, Huanyu Yu, Xiaokang Yang, and Chen Yao. 2018. Video summarization via semantic attended networks. In *AAAI*.

Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. 2019. Fine-grained action retrieval through multiple parts-of-speech embeddings. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 450–459.

Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. 2019. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6602–6611.

Yuxiang Wu and Baotian Hu. 2018. Learning to extract coherent summary via deep reinforcement learning. In *AAAI*, pages 5602–5609.

Shuwen Xiao, Zhou Zhao, Zijian Zhang, Xiaohui Yan, and Min Yang. 2020. Convolutional hierarchical attention network for query-focused video summarization. *arXiv preprint arXiv:2002.03740*.

Yuexiang Xie, Fei Sun, Yang Deng, Yaliang Li, and Bolin Ding. 2021. Factual consistency evaluation for text summarization via counterfactual estimation. In *Conference on Empirical Methods in Natural Language Processing*.

Ke Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. 2021. Taco: Token-aware cascade contrastive learning for video-text alignment. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11542–11552.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *ECCV*.

Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. 2017. End-to-end concept word detection for video captioning, retrieval, and question answering. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3261–3269.

S. Yuan, K. Bai, Liqun Chen, Yizhe Zhang, Chenyang Tao, C. Li, Guoyin Wang, R. Henao, and L. Carin. 2020. Weakly supervised cross-domain alignment with optimal transport. *BMVC*.

Yitian Yuan, Tao Mei, Peng Cui, and Wenwu Zhu. 2019. Video summarization by learning deep side semantic embedding. *IEEE Transactions on Circuits and Systems for Video Technology*, 29:226–237.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark. Association for Computational Linguistics.

Bowen Zhang, Hexiang Hu, and Fei Sha. 2018. Cross-modal and hierarchical modeling of video and text. In *ECCV*.

Haoxin Zhang, Zhimin Li, and Qinglin Lu. 2021. Better learning shot boundary detection via multi-task. *Proceedings of the 29th ACM International Conference on Multimedia*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *ArXiv*, abs/1912.08777.

Litian Zhang, Xiaoming Zhang, Junshu Pan, and Feiran Huang. 2022. Hierarchical cross-modality semantic correlation learning model for multimodal summarization. In *AAAI*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019b. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019c. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5059–5069, Florence, Italy. Association for Computational Linguistics.

Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. 2017. Temporal action detection with structured segment networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2933–2942.

Feng Zhou, Fernando De la Torre, and Jessica K. Hodgins. 2013. Hierarchical aligned cluster analysis for temporal clustering of human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:582–596.

Kaiyang Zhou, Yu Qiao, and Tao Xiang. 2018a. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *AAAI*, pages 7582–7589.

Kaiyang Zhou, T. Xiang, and A. Cavallaro. 2018b. Video summarisation by classification with deep reinforcement learning. In *BMVC*.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018c. Multi-turn response selection for chatbots with deep attention matching network. In *ACL*.

Jiacheng Zhu, Aritra Guha, Mengdi Xu, Yingchen Ma, Rayleigh Lei, Vincenzo Loffredo, XuanLong Nguyen, and Ding Zhao. 2021. Functional optimal transport: Mapping estimation and domain adaptation for functional data. *ArXiv*, abs/2102.03895.

Junnan Zhu, Haoran Li, Tianshan Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. Msmo: Multimodal summarization with multimodal output. In *EMNLP*.

Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020. Multimodal summarization with guidance of multimodal reference. In *AAAI*.

A Optimal Transport (OT) Basis

OT is the problem of transporting mass between two discrete distributions supported on latent feature space $\mathcal{X}$. Let $\boldsymbol{\mu} = \{\boldsymbol{x}_i, \boldsymbol{\mu}_i\}_{i=1}^{n}$ and $\boldsymbol{v} = \{\boldsymbol{y}_j, \boldsymbol{v}_j\}_{j=1}^{m}$ be the discrete distributions of interest, where $\boldsymbol{x}_i, \boldsymbol{y}_j \in \mathcal{X}$ denotes the spatial locations and μ_i, v_j, respectively, denoting the non-negative masses. Without loss of generality, we assume $\sum_i \mu_i = \sum_j v_j = 1$. $\pi \in \mathbb{R}_{+}^{n \times m}$ is a valid transport plan if its row and column marginals match μ and $\boldsymbol{v}$, respectively, which is $\sum_i \pi_{ij} = v_j$ and $\sum_j \pi_{ij} = \mu_i$. Intuitively, π transports π_{ij} units of mass at location $\boldsymbol{x}_i$ to new location $\boldsymbol{y}_j$. Such transport plans are not unique, and one often seeks a solution $\pi^* \in \Pi(\boldsymbol{\mu}, \boldsymbol{v})$ that is most preferable in other ways, where $\Pi(\boldsymbol{\mu}, \boldsymbol{v})$ denotes the set of all viable transport plans. OT finds a solution that is most cost effective w.r.t. cost function $C(\boldsymbol{x}, \boldsymbol{y})$:

$$\mathcal{D}(\boldsymbol{\mu}, \boldsymbol{v}) = \sum_{ij} \pi_{ij}^{*} C\left(\boldsymbol{x}_i, \boldsymbol{y}_j\right) = \inf_{\pi \in \Pi(\mu, v)} \sum_{ij} \pi_{ij} C\left(\boldsymbol{x}_i, \boldsymbol{y}_j\right)$$

where $\mathcal{D}(\boldsymbol{\mu}, \boldsymbol{v})$ is known as OT distance. $\mathcal{D}(\boldsymbol{\mu}, \boldsymbol{v})$ minimizes the transport cost from $\boldsymbol{\mu}$ to $\boldsymbol{v}$ w.r.t. $C(\boldsymbol{x}, \boldsymbol{y})$. When $C(\boldsymbol{x}, \boldsymbol{y})$ defines a distance metric on $\mathcal{X}$, and $\mathcal{D}(\boldsymbol{\mu}, \boldsymbol{v})$ induces a distance metric on the space of probability distributions supported on $\mathcal{X}$, it becomes the Wasserstein Distance (WD).

B More Related Work

Optimal Transport OT studies the geometry of probability spaces (Villani, 2003), a formalism for finding and quantifying mass movement from one probability distribution to another. OT defines the Wasserstein metric between probability distributions, revealing a canonical geometric structure with rich properties to be exploited. The earliest contribution to OT originated from Monge in the eighteenth century. Kantorovich rediscovered it under a different formalism, namely the Linear Programming formulation of OT. With the development of scalable solvers, OT is widely applied to many real-world problems and applications (Flamary et al., 2021; Chen et al., 2020a; Yuan et al., 2020; Zhu et al., 2021; Klicpera et al., 2021; Alqahtani et al., 2021; Lee et al., 2019; Chen et al., 2019; Duan et al., 2022).

Video Summarization Video summarization aims at generating a short synopsis that summarizes the video content by selecting the most informative and vital parts. The summary usually contains a set of representative video keyframes or video key-fragments that have been stitched in chronological order to form a shorter video. The former type is known as video storyboard, and the latter one is known as video skim (Apostolidis et al., 2021). Traditional video summarization methods only use visual information, extracting important frames to represent the video content. For instance, Gygli et al. (2014); Jadon and Jasim (2020) generated video summaries by selecting keyframes using SumMe and TVSum datasets. Some category-driven or supervised training approaches were proposed to generate video summaries with video-level labels (Song et al., 2015; Zhou et al., 2018a; Xiao et al., 2020; Zhou et al., 2018b).

Textual Summarization Textual summarization takes textual metadata, i.e., documents, articles, tweets, etc, as input, and generates textual summaries, in two directions: abstractive summarization and extractive summarization. Abstractive methods select words based on semantic understanding, and even the words may not appear in the source (Tan et al., 2017; See et al., 2017b). Extractive methods attempt to summarize language by selecting a subset of words that retain the most critical points, which weights the essential part of sentences to form the summary (Narayan et al., 2018; Wu and Hu, 2018). Recently, the fine-tuning approaches have improved the quality of generated summaries based on pre-trained language models in a wide range of tasks (Liu and Lapata, 2019; Zhang et al., 2019c).

Video Temporal Segmentation Video temporal segmentation aims at generating small video segments based on the content or topics of the video, which is a fundamental step in content-based video analysis and plays a crucial role in video analysis. Previous work mostly formed a classification problem to detect the segment boundaries in the supervised manner (Sidiropoulos et al., 2011; Zhou et al., 2013; Poleg et al., 2014; Sokeh et al., 2018; Aakur and Sarkar, 2019). Recently, unsupervised methods have also been explored (Gygli et al., 2014; Song et al., 2015). Temporal segmentation of actions in videos has also been widely explored in previous works (Wang et al., 2019; Zhao et al., 2017; Lea et al., 2017; Kuehne et al., 2020; Sarfraz et al., 2021; Wang et al., 2020b). Video shot boundary detection and scene detection tasks are also relevant and has been explored in many previous studies

(Hassanien et al., 2017; Hato and Abdulmunem, 2019; Rao et al., 2020; Chen et al., 2021; Zhang et al., 2021), which aim at finding the visual change or scene boundaries.

Textual Segmentation Textual segmentation aim at dividing the text into coherent, contiguous, and semantically meaningful segments (Nicholls, 2021). These segments can be composed of words, sentences, or topics, where the types of text include blogs, articles, news, video transcript, etc. Previous work focused on heuristics-based methods (Koshorek et al., 2018; Choi, 2000), LDA-based modeling algorithms (Blei et al., 2003; Chen et al., 2009), or Bayesian methods (Chen et al., 2009; Riedl and Biemann, 2012). Recent developments in NLP developed large models to learn huge amount of data in the supervised manner (Mikolov et al., 2013; Pennington et al., 2014; Li et al., 2018; Wang et al., 2018). Besides, unsupervised or weakly-supervised methods has also drawn much attention (Glavas et al., 2016; Lukasik et al., 2020).

C Baselines

C.1 Baselines for the VMSMO dataset

For the VMSMO dataset, we compare with multimodal summarization baselines and textual summarization baselines:

Multimodal summarization baselines:

Synergistic (Guo et al., 2019): Guo et al. (2019) proposed a image-question-answer synergistic network to value the role of the answer for precise visual dialog, which is able to jointly learn the representation of the image, question, answer, and history in a single step.

PSAC (Li et al., 2019): The Positional Self-Attention with Coattention (PSAC) model adopted positional self-attention block to model the data dependencies and video-question co-attention to help attend to both visual and textual information.

MSMO (Zhu et al., 2018): MSMO was the first model on producing multimodal output as summarization results, which adopted the pointer-generator network, added attention to text and images when generating textual summary, and used visual coverage by the sum of visual attention distributions to select pictures.

MOF (Zhu et al., 2020): Zhu et al. (2020) proposed a multimodal objective function with the guidance of multimodal reference to use the loss from the summary generation and the image selection to solve the modality-bias problem.

DIMS (Mingzhe et al., 2020): DIMS is a dual interaction module and multimodal generator, where conditional self-attention mechanism is used to capture local semantic information within video, and the global-attention mechanism is applied to handle the semantic relationship between news text and video from a high level.

Textual summarization baselines:

Lead (Nallapati et al., 2017): The Lead method simply selects the first sentence of article/document as the textual summary.

TexkRank (Mihalcea and Tarau, 2004): TexkRank is a graph-based extractive summarization method which adds sentences as nodes and uses edges to weight similarity.

PG (See et al., 2017b): PG is a hybrid pointer-generator model with coverage, which copied words via pointing, and generated words from a fixed vocabulary with attention.

Unified (Hsu et al., 2018): The Unified model combined the strength of extractive and abstractive summarization, where a sentence-level attention is used to modulate the word-level attention and an inconsistency loss function is introduced to penalize the inconsistency between two levels of attentions.

GPG (Shen et al., 2019): Generalized Pointer Generator (GPG) replaced the hard copy component with a more general soft "editing" function, which learns a relation embedding to transform the pointed word into a target embedding.

C.2 Baselines for the Daily Mail and CNN datasets

For Daily Mail and CNN datasets, we have multimodal baselines, video summarization baselines, and textual summarization baselines:

Multimodal summarization baselines:

MSMO (Zhu et al., 2018): MSMO was the first model on producing multimodal output as summarization results, which adopted the pointer-generator network, added attention to text and images when generating textual summary, and used visual coverage by the sum of visual attention distributions to select pictures.

Img+Trans (Hori et al., 2019): (Hori et al., 2019) applied multi-modal video features including video frames, transcripts, and dialog context for dialog generation.

TFN (Zadeh et al., 2017): Tensor Fusion Network (TFN) models intra-modality and inter-modality dynamics for multimodal sentiment analysis which explicitly represents unimodal, bimodal, and trimodal interactions between behaviors.

HNNattTI (Chen and Zhuge, 2018): HNNattTI aligned the sentences and accompanying images by using attention mechanism.

M^{2}SM (Fu et al., 2021, 2020): M^{2}SM is a multimodal summarization model with a bi-stream summarization strategy for training by sharing the ability to refine significant information from long materials in text and video summarization.

Video summarization baselines:

VSUMM (De Avila et al., 2011): VSUMM is a methodology for the production of static video summaries, which extracted color features from video frames and adopted k-means for clustering.

DR-DSN (Zhou et al., 2018a): Zhou et al. (2018a) formulated video summarization as a sequential decision making process and developed a deep summarization network (DSN) to summarize videos. DSN predicted a probability for each frame, which indicates the likelihood of a frame being selected, and then takes actions based on the probability distributions to select frames to from video summaries.

Textual summarization baselines:

Lead3 (See et al., 2017a): Similar to Lead, Lead3 means picking the first three sentences as the summary result.

NN-SE (Cheng and Lapata, 2016): NN-SE is a general framework for single-document summarization composed of a hierarchical document encoder and an attention-based extractor.

T5 (Raffel et al., 2019): T5 is an encoder-decoder model pre-trained on a multi-task mixture of unsupervised and supervised tasks and for which each task is converted into a text-to-text format. T5 works well on a variety of tasks out-of-the-box by prepending a different prefix to the input corresponding to each task, inluding summarization.

Pegasus (Zhang et al., 2019a): Pre-training with Extracted Gap-sentences for Abstractive SUmmarization Sequence-to-sequence models (PEGASUS) uses self-supervised objective Gap Sentences Generation (GSG) to train a transformer encoder-decoder model.

BART (Lewis et al., 2020): BART is a sequence-to-sequence model trained as a denoising autoencoder, and showed great performance a variety of text summarization datasets.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7

☒ A2. Did you discuss any potential risks of your work?
To the best of our knowledge, we do not foresee any harmful uses of this study

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 4

☑ B1. Did you cite the creators of artifacts you used?
Section 4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 4

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 4

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Not applicable. Left blank.

C ☑ Did you run computational experiments?

Section 4 and Section 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4 and Section 5

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4 and Section 5

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4 and Section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4 and Section 5

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.