Towards Unified Uni- and Multi-modal News Headline Generation

Mateusz Krubiński and **Pavel Pecina**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
{krubinski,pecina}@ufal.mff.cuni.cz

Abstract

Thanks to the recent progress in vision-language modeling and the evolving nature of news consumption, the tasks of automatic summarization and headline generation based on multimodal news articles have been gaining popularity. One of the limitations of the current approaches is caused by the commonly used sophisticated modular architectures built upon hierarchical cross-modal encoders and modality-specific decoders, which restrict the model's applicability to specific data modalities – once trained on, e.g., *text+video* pairs there is no straightforward way to apply the model to *text+image* or *text-only* data. In this work, we propose a unified task formulation that utilizes a simple encoder-decoder model to generate headlines from uni- and multi-modal news articles. This model is trained jointly on data of several modalities and extends the textual decoder to handle the multimodal output.

1 Introduction

The task of Multimodal Summarization was introduced as an extension of the traditional NLP task of Text Summarization. Early works (e.g., Li et al., 2017; Sanabria et al., 2018; Li et al., 2020a) explored to what extent enriching the textual document with additional context-specific information (e.g., visual clues from images attached to a product/service review or video clips attached to a cooking recipe) helps the automatic systems in refining the summary generation process. Zhu et al. (2018) were the first to notice that the *informativeness* of a summary can be significantly improved by including the visual clues in the output, introducing the task of Multimodal Summarization with Multimodal Output (MSMO). In their formulation, based on a textual document and a set of images, the model is tasked to generate the textual summary and pick a single image as the *pictorial summary*. Li et al. (2020b) introduced a variant of the task where the input is a pair of textual article and a short video. The following works (e.g., Qiu et al., 2022; Zhang et al., 2023b) explored the challenging problem of multi-modal fusion and alignment by introducing auxiliary tasks during training and extending the model architecture with task-specific blocks. However, by doing so, the model is tailored to a specific data modality.

In this work, we propose a novel MSMO task formulation that supports the most common data modalities (*text+video→text+image*, *text+images→text+image*, *text→text*) with a single sequence-to-sequence model (Section 2). We explore two approaches (Section 3.2): i) extending a text-to-text baseline with visual features and ii) fine-tuning a multimodal foundation model. We show that the proposed unified formulation leads to results competitive with previously introduced task-specific solutions (Section 4) while not being restricted to specific data modalities.

2 Unifying MSMO

Previous works explored two variants of the MSMO task: video-based and image-based. In the video-based one, the multimodal article is represented as a pair of a video clip and a textual document. The goal is to generate the textual summary and to choose a single frame that acts as a pictorial summary. In the image-based variant, the input is a *set* of images, i.e., there is no temporal dependency. The second difference comes from the ground truth image: in the image-based variant, we assume that the target is one of the input images. In the video-based one, there is no such assumption[1] – a similarity function is utilized to obtain the per-frame labels for training using the *most similar* one as a positive target. Our goal is to train a system

[1]The target image is often created by applying minimal edits, such as cropping or watermark removal. In addition, computational reasons require to down-sample the input frames, potentially dropping the *exact* one that is used as a target.

Findings of the Association for Computational Linguistics: EACL 2024, pages 437–450

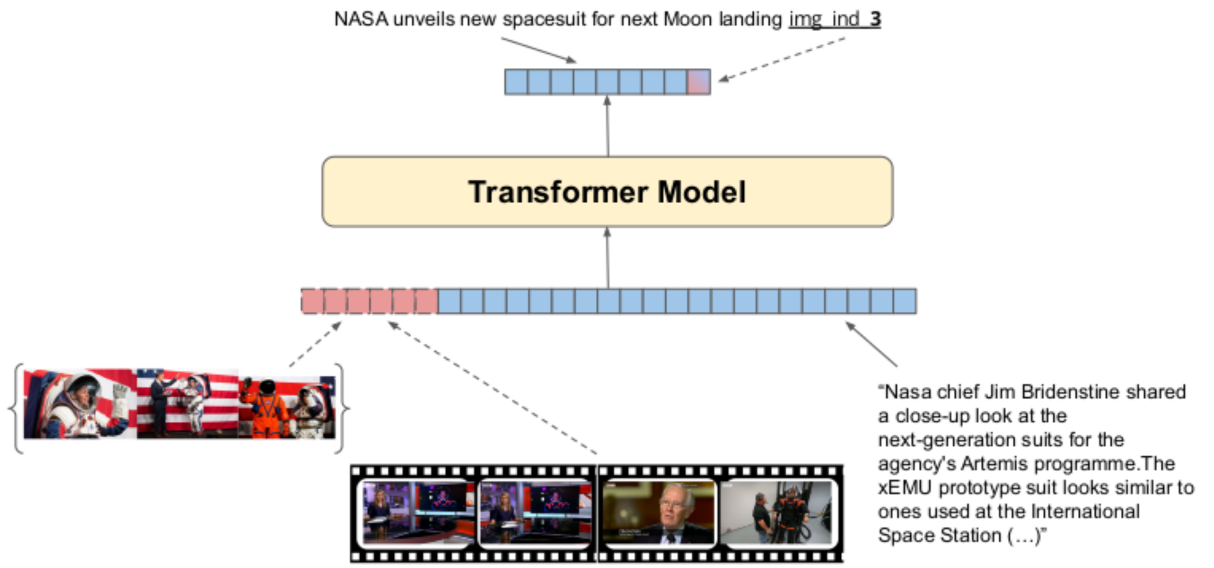

Figure 1: Overview of the proposed unified approach to MSMO. The visual tokens are appended to the text representation. The generated output includes the textual summary and the *index token* that indicates which input image (first, second, third, etc.) is picked as the pictorial summary. During training, a mixture of video-based, image-based, and text-only data is used.

capable of *natively* handling both MSMO variants as well as the basic text-to-text problem (summarization or headline generation). We achieve that by transforming the visual inputs into a sequence of image features that are concatenated with the textual token embeddings.

Instead of using a dedicated module for image scoring, we realize the target image representations by appending an *index token* to the textual target – `img_ind_1` indicates that the *first* image is the target, `img_ind_2` that the *second*, etc. This formulation allows us to use the standard Transformer architecture (Vaswani et al., 2017) trained end-to-end in a multi-task setting (see Figure 1) – for the text-only input, we do not extend the textual embeddings and do not add the index token into the target sequence.

3 Experiments

3.1 Data

In our experiments, we use the text-only `PENS` (Ao et al., 2021) dataset and the video-based `MLASK` (Krubiński and Pecina, 2023) dataset for training and testing. Since the largest publicly available image-based multimodal summarization dataset `M3LS` (Verma et al., 2023) lacks the image targets, we extend the English subset of the `M3LS` dataset by collecting the cover pictures on our own (see Appendix A for details). For brevity, we follow the TL;DW formulation by Tang et al. (2023) and use the article title as the textual target (i.e., the headline), although the proposed methods can also be applied for other summarization tasks, such as abstract generation.

3.2 Implementation

We use the T5 (Raffel et al., 2020) v1.1 base variant (250M trainable parameters) that we enrich with visual features extracted with frozen ViT-L/14 CLIP (Radford et al., 2021), projected with a linear layer to match the hidden dimension size (we refer to this model as T5CLIP). We extract a single vector per image (frame) and, following Wang et al. (2022a), use positional embeddings to indicate the temporal dimension for videos. We extend the model vocabulary with index tokens, i.e., «`img_ind_1`, `img_ind_2`, . . . » that are used for image/frame selection. We train with the Adafactor (Shazeer and Stern, 2018) optimizer using the default parameters from the Transformers (Wolf et al., 2020) package. For the multimodal baseline, we use the Flan T5-XL (Chung et al., 2023) version of BLIP-2 (Li et al., 2023, 3.9B parameters), which we extend to handle multiple images in the input – we concatenate the Q-Former features from multiple images before appending them to the textual embeddings introducing no new parameters. We use the LoRA (Hu et al., 2022) procedure and update only the `Q` and `V` matrices in the Q-Former

	ROUGE-L						BERTScore					
	MLASK		PENS		M3LS		MLASK		PENS		M3LS	
	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test
Lead	12.28	12.19	16.51	16.27	9.74	9.85	10.67	10.77	8.85	9.10	9.57	10.03
Oracle	**24.44**	**25.01**	38.99	39.17	23.85	23.65	21.09	21.99	31.78	31.91	18.43	19.34
Alpaca	14.81	15.07	26.80	26.92	16.54	16.96	18.67	19.14	28.40	28.62	19.34	20.78
BRIO	15.56	15.58	16.40	16.55	18.18	18.79	15.97	16.49	16.61	16.83	23.30	25.03
$\text{T5CLIP}_{\text{MLASK}}$	20.79	21.32	-	-	-	-	25.46	25.99	-	-	-	-
$\text{T5CLIP}_{\text{PENS}}$	-	-	43.00	44.21	-	-	-	-	**45.12**	**46.70**	-	-
$\text{T5CLIP}_{\text{M3LS}}$	-	-	-	-	29.63	**29.68**	-	-	-	-	**33.84**	**34.48**
T5CLIP	**21.48**	21.43	**43.07**	**44.47**	**29.64**	29.38	**26.43**	**26.36**	**45.24**	**46.80**	33.16	33.73
$\text{T5CLIP}_{w=10}$	**21.48**	**21.57**	42.60	43.74	29.32	29.28	25.98	**26.43**	44.31	45.74	32.67	33.25
$\text{T5CLIP}_{w=50}$	20.63	21.05	40.87	42.15	26.92	26.88	25.21	25.55	41.72	43.40	29.14	29.71
$\text{T5CLIP}_{\text{Smooth}}$	21.30	21.32	**43.25**	**44.39**	**30.06**	**30.03**	**26.50**	26.24	**45.53**	**46.94**	**33.70**	**34.44**
BLIP-2	**23.25**	**24.24**	**43.03**	**44.37**	**32.82**	**33.02**	**27.87**	**28.94**	44.56	46.27	**35.91**	**37.24**
MMS	19.99	20.07	-	-	-	-	23.97	24.38	-	-	-	-

Table 1: Evaluation of the textual output quality on the validation and test splits for each modality-specific dataset (Section 3.1). The three highest-scoring systems in each column are bolded independently for test-set and dev-set.

and Language Model components (5.7M trainable parameters), training with the AdamW (Loshchilov and Hutter, 2019) optimizer with β=(0.9, 0.999), learning rate of 1e-5 and weight decay of 5e-2. We train all the models for up to 10 epochs with early stopping applied if ROUGE-L F1 does not improve for 5 consecutive epochs. We limit the source size to 1024 sub-word tokens and the target length to 128 tokens. We train on a machine with three NVIDIA A40 GPUs and the average training time is 24 hours for the T5 variants (effective batch size 300) and one week for the BLIP-2 variant (effective batch size 60). During decoding, we utilize beam search of size 4, length penalty of 1.0, and repetition penalty (Keskar et al., 2019) of 2.5.

3.3 Metrics and baselines

Metrics We measure the quality of the textual output with ROUGE-L (Lin, 2004) and BERTScore (Zhang et al., 2020b), reporting the F1 scores. For the pictorial output, we report the cosine similarity (CosSim) between the ViT-L/14 CLIP features of the target image and the one chosen by the model. To measure the multi-modal interactions, we report the CLIPBERTScore (Wan and Bansal, 2022) metric. It is computed as a weighted average[2] of the CLIPScore (Hessel et al., 2021) of the chosen image and the generated summary and the BERTScore precision of the input article and the generated summary. For the image-based data, we also report the top-1 accuracy (Top-1 Acc), i.e., the percentage of predictions where the target image is correctly retrieved. For details, see Appendix B.

Baselines We report two extractive baselines: *Lead* that extracts the first sentence and *Oracle* that picks a sentence maximizing ROUGE-L with the ground truth. For the off-the-shelf textual abstractive baselines, we use the Alpaca (Taori et al., 2023) and BRIO (Liu et al., 2022) models (see Appendix C). For the video-based data, we compare with the MMS model (Krubiński and Pecina, 2023). We also report a trivial baseline *RandomVi* that picks a random image/frame. To further establish a comparison with the recent developments, we also report a generative visual baseline based on Stable Diffusion (Rombach et al., 2022). We employ the `stabilityai/stable-diffusion-2-1` model prompted with the textual target (`_TEXT_`) using the following template: "`High quality, photorealistic photo of _TEXT_`".

4 Results

Textual Output Table 1 compares the models (see examples of model outputs in Appendix D) trained separately on each task (e.g., $\text{T5CLIP}_{\text{PENS}}$) with the ones trained in the multi-task fashion (T5CLIP). The results are comparable, with additional textual data improving the performance on the smallest video-based dataset – MLASK. The proposed baselines, besides the *Oracle*, are lagging behind the task-specific models. The highest scores are obtained by the fine-tuned BLIP-2, which integrates the largest language component – Flan T5-XL.

[2] We use the recommended $\alpha = 0.25$

	CosSim				CLIPBERTScore				Top-1 Acc	
	MLASK		M3LS		MLASK		M3LS		M3LS	
	dev	test	dev	test	dev	test	dev	test	dev	test
RandomVi	0.61	0.61	0.75	0.76	-	-		-	33.20	33.59
T5CLIP$_{\text{MLASK}}$	0.64	0.64	-	-	70.56	70.59	-	-	-	-
T5CLIP$_{\text{M3LS}}$	-	-	**0.97**	**0.97**	-	-	69.57	69.70	**93.59**	**94.56**
T5CLIP	0.64	0.64	0.93	0.94	70.67	70.65	69.61	69.77	87.49	88.55
T5CLIP$_{w=10}$	0.64	0.64	0.96	**0.97**	70.99	70.99	69.74	69.92	93.03	94.05
T5CLIP$_{w=50}$	0.64	0.63	0.96	**0.97**	71.12	71.11	69.60	69.72	91.76	93.19
T5CLIP$_{\text{Smooth}}$	0.64	0.63	0.82	0.81	70.65	70.61	69.83	69.96	39.91	38.55
BLIP-2	0.63	0.62	0.83	0.84	71.46	71.44	**70.07**	**70.26**	60.46	61.73
MMS	**0.68**	**0.68**	-	-	**71.50**	**71.53**	-	-	-	-
Stable Diffusion v2.1	0.42	0.43	0.44	0.44	-	-	-	-	-	-

Table 2: Evaluation of the visual output quality on the validation and test splits for video-based and image-based datasets (Section 3.1). The highest-scoring system in each column is bolded independently for test-set and dev-set.

Visual Output The relatively high scores of the random visual baseline (Table 2) may indicate that the CLIP features are not distinctive enough for the closely related images/frames coming from the same article. The image-specific model (T5CLIP$_{\text{M3LS}}$) performs slightly better than the multi-task one (T5CLIP). We attribute this to the potentially easier image-based task formulation (Section 2) where the target input (i.e., one with CosSim $= 1.0$) is present in the input.

In order to improve the visual performance, we propose to use two methods: smooth labels (see Krubiński and Pecina, 2023) and greater weights w for the visual tokens when computing loss. Using 10 times greater weight (T5CLIP$_{\text{w=10}}$) improves the top-1 accuracy on M3LS, while using 50 times greater weight (T5CLIP$_{\text{w=50}}$) brings no further improvement, degrading the quality of textual output. The smooth labels (T5CLIP$_{\text{Smooth}}$), designed for video-based data, are not effective on image-based data. The highest similarities on MLASK are achieved by the MMS model, which uses a separate visual encoder and frame-scoring module. The highest CLIPBERTScore is achieved by MMS on MLASK (the best visual output quality) and BLIP-2 on M3LS (the best textual model, a greater weight for the textual component). Masking the visual features with random noise has a negligible effect on the textual output (M3LS test 29.38→29.32), which we attribute to the "greedy learning" hypothesis by Wu et al. (2022), but drops the top-1 accuracy to chance level (M3LS test 88.55→37.9).

5 Related Work

Historically, for both the video-based (Li et al., 2020b) and the image-based (Zhu et al., 2018) MSMO, the attention mechanism (Bahdanau et al., 2015) was used to condition the encoded text representation on the visual information, which in the next step was passed to the autoregressive text decoder. Following works focused on improving the quality end efficiency of this process: Li et al. (2018) and Liu et al. (2020) focused on the filtering mechanism that would allow the model to attend only to chosen relevant features avoiding potential noise. Yu et al. (2021) and Qiao et al. (2022) worked on adapting strong pre-trained language models to the multimodal input. All of those works perturb the textual representation – the model is no longer capable of inference on text-only data. The reverse attention (*vision→text*) was used to condition the visual information on the text content. Using a learning signal from the pictorial target, the model was trained to produce image/frame-level scores.

A step towards simplifying these modular approaches was recently made by Jiang et al. (2023), who generate pseudo-captions for input images and then pick the image with the highest similarity between the caption and the generated textual summary, and He et al. (2023), who instead of using a textual decoder, predict sentence-level scores and extract top-k sentences as the textual summary. A one-for-all architectures unifying several vision-and-language tasks have also been explored in a wider context. Cho et al. (2021) introduce visual sentinel tokens corresponding to image regions, allowing them to realize Visual Grounding with a text-only decoder. The Task- and Modality-Agnostic OFA framework (Wang et al., 2022b) unifies the multi-modal and text-only tasks with a sequence-to-sequence Transformer. By design, it is however limited to tasks dealing with a single image, e.g., Image Captioning or Visual Question Answering, not supporting inputs containing multiple images or videos. A recent line of research on multimodal LLMs (Zhang et al., 2023a; Maaz et al., 2023; Li et al., 2024) transfers the knowledge from image-text models into video-text models.

Inspired by those works and the general-purpose multimodal foundation models (e.g., Bao et al., 2022; Alayrac et al., 2022; Wang et al., 2023a), we propose the unified formulation (Section 2) – the multi-task training with a simplified encoder allows the model to natively handle both multi-modal and text-only input and the usage of *index tokens* that explicitly point to a particular input image allows us to drop the scoring module and train with a single text decoder.

6 Conclusions

In this pilot study on multi-task multi-modal summarization, we propose a novel unified formulation for the MSMO task. By training the textual decoder to generate *index tokens*, we make use of the training signal from the visual modality without a dedicated scoring module. Our results indicate that multi-task training, which incorporates text-only data, is an alternative to text-only pretraining, which preserves the *native* capability to handle purely textual input. For the challenging task of video-based MSMO, there is still some gap left when it comes to the visual output quality when compared to sophisticated task-specific architecture. Based on our results, for this specific task, the visual generative approaches are still inferior to extractive ones.

Limitations

Multimodal Summarization variants. In our work, we examine three variants of the multimodal summarization task: *text+video→→text+image*, *text+images→text+image*, and *text→text*. We acknowledge existence of other formulations, such as *text+video→text* (Qiao et al., 2022), *images→text* (Trieu et al., 2020) or *video→text+images* (Lin et al., 2023) that we did not include in our experiments.

Dataset choice. Our findings are based on particular datasets, in a particular language (English) and from a particular domain (news articles). The fact that the previously introduced datasets (Li et al., 2020b; Tang et al., 2023) are not publicly available is a limiting factor.

Extension of the M3LS dataset. Since the largest image-based dataset (Section 3.1) lacks the cover pictures in the training data, we collected them by automatically crawling a news website. To check the validity of our setup, we sampled 100 articles and manually checked the collected images, but no large-scale human evaluation was conducted.

Generative models. Both of the off-the-shelf generative models that we use: the visual one (*Stable Diffusion v2-1*) and the textual one (*Alpaca*) were trained on data that potentially may include harmful content such as explicit pornographic materials or toxic, stereotyped language. We did not apply any filtering to the model outputs, so the predictions may not be free of bias.

Acknowledgements

This work was supported by the Czech Science Foundation (grant no. 19-26934X) and CELSA (project no. 19/018). In this work, we used data and tools provided by the LINDAT/CLARIAH-CZ Research Infrastructure (`https://lindat.cz`), supported by the Ministry of Education, Youth and Sports of the Czech Republic (project no. LM2023062).

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.

Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. 2021. PENS: A dataset and generic framework for personalized news headline generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 82–92, Online. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly

Learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022. VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts. In *Advances in Neural Information Processing Systems*, volume 35, pages 32897–32912. Curran Associates, Inc.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying Vision-and-Language Tasks via Text Generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1931–1942. PMLR.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2023. Scaling Instruction-Finetuned Language Models. *arXiv preprint arXiv:2210.11416*.

Bo He, Jun Wang, Jielin Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. 2023. Align and Attend: Multimodal Summarization with Dual Contrastive Losses. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 14867–14878. IEEE.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Chaoya Jiang, Rui Xie, Wei Ye, Jinan Sun, and Shikun Zhang. 2023. Exploiting pseudo image captions for multimodal summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 161–175, Toronto, Canada. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*.

Mateusz Krubiński and Pavel Pecina. 2023. MLASK: Multimodal summarization of video-based news articles. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 910–924, Dubrovnik, Croatia. Association for Computational Linguistics.

Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020a. Aspect-Aware Multimodal Summarization for Chinese E-Commerce Products. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8188–8195.

Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, and Chengqing Zong. 2018. Multi-modal Sentence Summarization with Modality Attention and Image Filtering. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4152–4158. International Joint Conferences on Artificial Intelligence Organization.

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 *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1092–1102, Copenhagen, Denmark. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2024. VideoChat: Chat-Centric Video Understanding. *arXiv preprint arXiv:2305.06355*.

Mingzhe Li, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2020b. VMSMO: Learning to generate multimodal summary for video-based news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9360–9369, Online. Association for Computational Linguistics.

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

Jingyang Lin, Hang Hua, Ming Chen, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Jiebo Luo. 2023. VideoXum: Cross-modal Visual and Textural Summarization of Videos. In *IEEE Transactions on Multimedia*, pages 1–13. IEEE.

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

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. BRIO: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. *arXiv preprint arXiv:2306.05424*.

Benjamin Minixhofer, Jonas Pfeiffer, and Ivan Vulić. 2023. Where's the point? self-supervised multilingual punctuation-agnostic sentence segmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7215–7235, Toronto, Canada. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11(4381):1–15.

Lingfeng Qiao, Chen Wu, Ye Liu, Haoyuan Peng, Di Yin, and Bo Ren. 2022. Grafting pre-trained models for multimodal headline generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 244–253, Abu Dhabi, UAE. Association for Computational Linguistics.

Jielin Qiu, Jiacheng Zhu, Mengdi Xu, Franck Dernoncourt, Trung Bui, Zhaowen Wang, Bo Li, Ding Zhao, and Hailin Jin. 2022. MHMS: Multimodal Hierarchical Multimedia Summarization. *arXiv preprint arXiv:2204.03734*.

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 *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A Large-scale Dataset for Multimodal Language Understanding. In *Proceedings of the Workshop on Visually Grounded Interaction and Language (NeurIPS 2018)*.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive Learning Rates with Sublinear Memory Cost. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4596–4604. PMLR.

Peggy Tang, Kun Hu, Lei Zhang, Jiebo Luo, and Zhiyong Wang. 2023. TLDW: Extreme Multimodal Summarisation of News Videos. In *IEEE Transactions on Circuits and Systems for Video Technology*. IEEE.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

Ni Trieu, Sebastian Goodman, P. Narayana, Kazoo Sone, and Radu Soricut. 2020. Multi-Image Summarization: Textual Summary from a Set of Cohesive Images. *arXiv preprint arXiv:2006.08686*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz

Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yash Verma, Anubhav Jangra, Raghvendra Verma, and Sriparna Saha. 2023. Large scale multi-lingual multi-modal summarization dataset. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3620–3632, Dubrovnik, Croatia. Association for Computational Linguistics.

David Wan and Mohit Bansal. 2022. Evaluating and improving factuality in multimodal abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9632–9648, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022a. GIT: A Generative Image-to-text Transformer for Vision and Language. *Transactions on Machine Learning Research*.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022b. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 2023a. Image as a Foreign Language: BEiT Pretraining for Vision and Vision-Language Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19175–19186. IEEE.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J. Geras. 2022. Characterizing and Overcoming the Greedy Nature of Learning in Multi-modal Deep Neural Networks. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 24043–24055. PMLR.

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

Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553, Singapore. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Litian Zhang, Xiaoming Zhang, Ziming Guo, and Zhipeng Liu. 2023b. CISum: Learning Cross-modality Interaction to Enhance Multimodal Semantic Coverage for Multimodal Summarization. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, pages 370–378.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. MSMO: Multimodal summarization with multimodal output. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4154–4164, Brussels, Belgium. Association for Computational Linguistics.

A Appendix – Data preparation

A.1 MLASK

Since the textual part of MLASK[3] – the largest publicly available video-based news summarization dataset – is in the Czech language, we used the CUBBITT (Popel et al., 2020) Machine Translation system[4] to translate articles and summaries (titles) into English. We use the split proposed by Krubiński and Pecina (2023), i.e., 36,109/2,482/2,652 instances for training/validation/testing. In our early experiments, we sampled one of every 25 frames (1 frame per second), which on average produced 86 images (frames) per video, with the longest videos having up to about 300 frames sampled. This number is too large to process with the BLIP-2 model – it uses the Q-Former to map each input image into 32 visual tokens, which would require us to process sequences of length up to 9,600. Therefore, we decided to further down-sample the input by sampling 20 frames evenly spaced across the video. To check whether this affects the model performance, we trained the $\text{T5CLIP}_{\text{MLASK ALL}}$ variant (see Section 3.2) that uses the denser sampling for each video. The results (MLASK dev-set ROUGE-L: 20.79 $\rightarrow$ 20.55, BERTScore: 25.46 $\rightarrow$ 25.34, CosSim: 0.64 $\rightarrow$ 0.61) indicate that the model is not able to make use of the dense frame sampling, showing that the problem of frame-selection requires more work in the future.

A.2 PENS

The PENS dataset[5] contains 113,762 news articles and was originally introduced for personalized news headline generation. We filtered it by removing articles identified as non-English by the langid[6] language identifier, and those where the title has less than 2 words or more than 25 words. In the next step, we de-duplicated the data based on the article and title fields. We were left with 100,992 documents (89%), out of which 5,000 were used for validation and testing and the remaining ones (90,992) for training.

A.3 M3LS

The M3LS dataset[7] was introduced recently as the largest resource for image-based multimodal summarization. The data was collected in several languages, including 376,367 documents in English, from the `www.bbc.com/news` website. However, the multimodal information (images) is present only on the source side – the target is purely textual. In order to extend this resource with the visual target, we made use of the URLs that were provided for each article by collecting the content (URL) of the meta element HTML tag with `property="og:image"`. Based on our understanding and manual checks, the URLs correspond to the picture that is used to visually represent the article at the `www.bbc.com/news` main page. In the next step, we collected the images and applied two-step filtering: we kept only those images that had a particular resolution (1024x490), and in the next step, we removed duplicates. Finally, we filtered those multimodal articles that fulfilled two conditions: they had at least a single image in the input and we were able to collect the target image for them. We ended up with 115,432 instances, which we split into training/validation/testing based on the publication date: articles published in January–April of 2021 for validation (5,865 instances) and the ones published in May–October of 2021 for testing (6,854 instances). The remaining data (before January 2021) is used for training (102,713 instances). Following the image-based MSMO formulation (Section 2), we append the target image to the source images, shuffling them during training to avoid positional bias. The quantitative statistics of the number of input images in the extended M3LS dataset are displayed in Table 3.

Min	Q_1	Mean	Q_3	Max
2	2	3.79	4	21

Table 3: Quantitative statistics of the number of input images (including the target image) in the subset of the English M3LS dataset that we extended with the multimodal target.

[3] https://github.com/ufal/MLASK
[4] https://ufal.mff.cuni.cz/cubbitt
[5] https://msnews.github.io/pens_data.html
[6] https://github.com/saffsd/langid.py
[7] https://github.com/Raghvendra-14/M3LS

B Appendix – Metrics

We use the ROUGE metric from the TorchMetrics package[8] and the original implementations of BERTScore[9] and CLIPBERTScore[10]. The signature of the BERTScore model that we use is: `roberta-large_L17_no-idf_version0̄.3.12(hug_trans=4.29.0.dev0)-rescaled`. For readability reasons, we re-scale both BERTScore and CLIPBERTScore into the [0–100] range by multiplying the numerical scores by 100.

C Appendix – Baselines

The Stanford Alpaca model[11] is a text-only, Transformer-based Large Language Model (LLM), fine-tuned from the LLaMA (Touvron et al., 2023) model to follow instructions. It has been trained on the automatically generated data created with the Self-Instruct (Wang et al., 2023b) techniques. In our experiments, we use the following prompt:

```
Below is an instruction that describes a task, paired with an input that provides
further context. Write a response that appropriately completes the request.

### Instruction:
Generate a one sentence summary of a given text, using no more than 10 words.

### Input:
__DOCUMENT_TEXT__

### Response:"
```

We report results with the 7B parameter variant and, for generation, utilize beam search of size 4, length penalty of -5.0, and repetition penalty of 2.5. In our early experiments, we noticed that truncating the input at the token level resulted in words and sentences being cut in half, which negatively affected the model performance. To avoid this, we use the `wtpsplit` package (Minixhofer et al., 2023) to prompt the model with full sentences, capping the input length (i.e., `__DOCUMENT_TEXT__`) at 1000 characters.

BRIO (Liu et al., 2022) is a recent encoder-decoder model trained for both summary *generation* and *evaluation*, i.e., the ability to score the quality of candidate summaries. We use the `Yale-LILY/brio-xsum-cased` variant (568M parameters), which is based upon the pre-trained PEGASUS (Zhang et al., 2020a) model and fine-tuned on the XSum (Narayan et al., 2018) dataset to generate single-sentence summaries.

When generating images with the `stabilityai/stable-diffusion-2-1` model, we use the standard inference parameters (`guidance_scale=5` and `num_inference_steps=50`) with the following `negative_prompt`: *"ugly, tiling, poorly drawn hands, poorly drawn feet, poorly drawn face, out of frame, extra limbs, disfigured, deformed, body out of frame, bad anatomy, watermark, signature, cut off, low contrast, underexposed, overexposed, bad art, beginner, amateur, distorted face"*.

[8] https://torchmetrics.readthedocs.io/en/stable/text/rouge_score.html
[9] https://github.com/Tiiiger/bert_score
[10] https://github.com/meetdavidwan/faithful-multimodal-summ
[11] https://github.com/tatsu-lab/stanford_alpaca

D Appendix – Model Outputs

```
Walrus counting from space: How many tusked
beasts do you see?
```

(a) **Reference**

```
Thousands of volunteers to count Arctic
walruses from space
```

(b) **T5CLIP**

```
Scientists count walruses from space
```

(c) **BLIP-2**

```
Walruses are heavily dependent on sea-ice,
which has been in sharp retreat, leading
to increased difficulty for the animals to
hunt and rest.
```

(d) **Stable Diffusion 2.1 + Alpaca**

Figure 2: Pictorial summary – M3LS Example 1.

Ireland’s Naval Service marks its 75th anniversary

(a) **Reference**

’I thought the navy was cool and really interesting’

(b) **T5CLIP**

Irish Navy celebrates 75th anniversary

(c) **BLIP-2**

Covid has ensured that anniversary commemorations will be more subdued than the 50th anniversary celebrations, when foreign navies visited Ireland.

(d) **Stable Diffusion 2.1 + Alpaca**

Figure 3: Pictorial summary – M3LS Example 2.

Man seriously injured his head at waste treatment company, helicopter flew for him

(a) **Reference**

A worker was injured in a truck at a waste treatment plant in Prague

(b) **T5CLIP**

A man was injured at a waste treatment company in Prague. He was airlifted to hospital

(c) **BLIP-2**

A man was injured in a waste treatment company in Prague. He died at the scene

(d) **MMS**

Man injured at waste treatment plant, airlifted conscious to hospital.

(e) **Stable Diffusion 2.1 + Alpaca**

Figure 4: Pictorial summary – MLASK Example 1.

I will make the universe accessible to all of you, exulted Branson. Prepare 5 million

(a) **Reference**

Branson's "a once-in-a-lifetime experience". Take a ride in space with his crew

(b) **T5CLIP**

Richard Branson became the second 70-year-old to go into space

(c) **BLIP-2**

The world's richest man has a new era of space travel, Branson and his family are heading to the edge of space

(d) **MMS**

Virgin Galactic successfully completed its first commercial space flight, marking a major milestone for space tourism.

(e) **Stable Diffusion 2.1 + Alpaca**

Figure 5: Pictorial summary – MLASK Example 2.