MLASK: Multimodal Summarization of Video-based News Articles

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

In recent years, the pattern of news consumption has been changing. The most popular multimedia news formats are now multimodal – the reader is often presented not only with a textual article but also with a short, vivid video. To draw the attention of the reader, such video-based articles are usually presented as a short textual summary paired with an image thumbnail. In this paper, we introduce MLASK[1] (**M**ultimoda**L** **A**rticle **S**ummarization **K**it) – a new dataset of video-based news articles paired with a textual summary and a cover picture, all obtained by automatically crawling several news websites. We demonstrate how the proposed dataset can be used to model the task of multimodal summarization by training a Transformer-based neural model. We also examine the effects of pre-training when the usage of generative pre-trained language models helps to improve the model performance, but (additional) pre-training on the simpler task of text summarization yields even better results. Our experiments suggest that the benefits of pre-training and using additional modalities in the input are not orthogonal.

1 Introduction

Automatic summarization is one of the basic tasks both in Natural Language Processing – text summarization – and in Computer Vision – video summarization. Multimodal summarization (MMS) builds a bridge between those two fields.

Early works on multimodal summarization explored the usage of the secondary modality as an auxiliary source of information to guide the refinement process of the main modality. Li et al. (2017) collected videos and news articles covering a handcrafted list of recent significant world events by querying a web search engine and trained a model to mimic the reference summaries written by human annotators. Zhu et al. (2018) were the first to introduce the task of Multimodal Summarization with Multimodal Output (MSMO). They collected a large-scale dataset of news articles paired with corresponding images and trained a system to generate a textual summary and choose a single image as a pictorial summary. By introducing the multimodal output, a uni-modal solution was no longer sufficient as a baseline. Building upon this, Li et al. (2020b) extended the task to video-based MSMO. Based on a textual document and a short video clip, besides generating the textual summary, the system was also challenged to select a single frame from the video as a cover picture.

We believe there is still a lot of questions that remain unanswered, e.g.: *How to evaluate multimodal outputs?* or *How to approach pre-training?* In this paper, we contribute to the area of video-based MSMO (VMSMO) by: **1)** introducing a full-scale VMSMO dataset in Czech, extending the very limited available resources for this task to a new language; **2)** exploring the pre-training strategies by transferring knowledge from the simpler task of text-to-text summarization; **3)** re-defining the training labels to consider intra-video similarities; **4)** proposing a human evaluation framework for assessing the quality of VMSMO.

2 Related Work

In our work, we build upon recent advances in three fields: text summarization, video summarization, and multimodal summarization.

2.1 Text Summarization

Text summarization aims to automatically produce a short fluent summary that preserves the crucial information from the source document(s). Historically, a majority of works focused on the news domain and English language (Nallapati et al., 2016; Grusky et al., 2018; Fabbri et al., 2019). Recently, new research directions, such as multilingual summarization (Scialom et al., 2020; Varab

[1] https://github.com/ufal/MLASK

Findings of the Association for Computational Linguistics: EACL 2023, pages 910–924

and Schluter, 2021) or dialogue summarization, (Gliwa et al., 2019; Zhong et al., 2021) have been explored. Fabbri et al. (2021) benchmarked over 20 recent summarization models and concluded that the abstractive summaries produced by pre-trained generative languages models fine-tuned on summarization datasets (Zhang et al., 2020; Lewis et al., 2020) consistently performed best with regards to both automatic metrics and human evaluation.

2.2 Video Summarization

Video summarization aims to refine the video content by either choosing a set of the most representative frames, known as a video storyboard, or selecting short video fragments, known as a video skim. As noted in the recent survey (Apostolidis et al., 2021), in both cases, the usual approach is to start with modeling the frame-level importance scores, which can then be aggregated to segment-level scores.

Contrary to text summarization, abstractive approaches that generate the summary from scratch are yet to be explored. The most relevant to our work are the recent publications on query-based video summarization, e.g., Li et al. (2023) and Huang et al. (2021), that use a text-based input to enrich frame-level representations and guide the summarization towards a user-specified query.

2.3 Multimodal Summarization

Previous works (e.g., Li et al., 2017, 2018; Palaskar et al., 2019) explored the addition of multimodal information such as video or audio transcript to enrich the textual document, aiming to generate better textual summaries. Zhu et al. (2018), who introduced the MSMO task, trained a model that jointly generated text and selected the most relevant image from a pre-defined set of images. Li et al. (2020b) and Fu et al. (2021) were the first to tackle the VMSMO problem. In their work, the cover picture choice was modeled as a frame selection problem. In the follow-up work (Tang et al., 2022), a video-article pair was summarized as a single frame and a one-sentence summary using an optimal transport-based unsupervised training strategy.

3 MLASK Dataset

Previous works on MMS operated on datasets in either English (Li et al., 2017, 2018; Palaskar et al., 2019; Fu et al., 2021; Tang et al., 2022) or Chinese (Li et al., 2020b; Li et al., 2020a). To extend the available resources, we collected a new dataset in a different language – Czech, a West Slavic language with a rich system of morphology and a relatively flexible word order.

	Mean	Q_1	Median	Q_3
Title	11.16 ± 2.78	9	11	13
Abstract	33.40 ± 13.86	22	32	43
Article	276.96 ± 191.74	154	231	343

Table 1: Quantitative statistics of the lengths of titles, abstracts, and full texts (measured in the number of tokens) for the MLASK dataset. Q_1 and Q_3 denote the first and the third quartile, respectively.

3.1 Data Preparation

The steps taken while preparing the dataset are:

1. Two Czech websites publishing news articles accompanied with a video clip, textual summary, and a cover picture were identified.
2. Based on the HTML structure of each website, the articles accompanied by a video clip (mp4) and a cover picture (jpeg) were downloaded.
3. From each relevant article, its title, abstract, and full text were extracted.
4. The following documents were dropped:
   - with videos longer than 5 minutes;
   - with full text shorter than 50 words or longer than 2,000 words;
   - with abstract shorter than 10 words or longer than 80 words;
   - with title shorter than 2 words;
   - with either the full text or abstract identified as non-Czech by the langid[2] language-identifier.
5. Every video was re-sampled to the same frame rate (25 fps) and resized to the same resolution (1280x720).

3.2 Dataset Size Statistics

In total, the collected dataset contains 41,243 instances, all including the article's text, title, abstract, video, and cover picture. The quantitative statistics of the data are displayed in Table 1. The average video duration is 85.58 seconds. For comparison, we also report the statistics of other datasets proposed for the VMSMO task so far (Table 2).

[2] https://github.com/saffsd/langid.py

Dataset	#Articles	Article Length	Summary Length	Video Length	Language
VMSMO (Li et al., 2020b)	184,920	97	11	60s	Chinese
MM-AVS (Fu et al., 2021)	2,173	685	57	109s	English
XMSMO-News (Tang et al., 2022)	4,891	102	12	346s	English
MLASK (this paper)	41,243	277	33	86s	Czech

Table 2: Comparison of the datasets introduced for the VMSMO task. The concrete statistics are reported as averages computed over the whole corpus. For the textual part, we report the average number of tokens.

4 Multimodal Summarization

In our experiments, a video-based news article is represented by a pair (V, X). V corresponds to the video input – a sequence of frames: $V = (v_1, v_2, \ldots, v_N)$. X is the news article presented as a sequence of tokens: $X = (x_1, x_2, \ldots, x_M)$. We assume that for each article, there is a ground-truth textual summary $Y = (y_1, y_2, \ldots, y_L)$ and a ground-truth cover picture P. The task is to generate a textual summary $\widehat{Y}$ that includes the main points of the article and to choose a frame $\widehat{v}$ to act as a cover picture (pictorial summary).

4.1 Overview

The proposed MMS model (see Figure 1) is structured into three parts: *Feature Encoder* composed of a text, video, and frame encoder, *Cross-modal Interaction Module* fusing the visual and textual representations, and *Multimodal Decoder* responsible for the summary generation and frame selection.

4.2 Feature Encoder

The Feature Encoder consists of a text encoder, video encoder, and frame encoder:

Text Encoder. We use the Transformer (Vaswani et al., 2017) encoder model to map the textual news article into the sequence of token embeddings (Eq. 1). Following the findings of Yu et al. (2021), we use the pre-trained mT5 model (Xue et al., 2021) to initialize its weights. We examine the influence of task-specific pre-training (Section 6.3) by fine-tuning the mT5 model on the simpler task of text-to-text summarization.

$$X_{enc} = \text{TransformerEncoder}(X) \quad (1)$$

Video Encoder. The news videos in our dataset are several minutes long and consist of hundreds of frames. To incorporate the short-term temporal dependencies, we employ the 3D convolutional networks. In our experiments, we segment the video into non-overlapping sequences of frames and use the 3D CNN network for feature extraction (Eq. 2). As the feature extractors, we use the R(2+1)D model trained by Ghadiyaram et al. (2019) for video action recognition on weakly-supervised social-media videos and the visual component of the S3D Text-Video model trained in a self-supervised manner by Miech et al. (2020) on the HowTo100M dataset (Miech et al., 2019). To incorporate the long-term temporal dependencies, we process the sequence of video features with the Transformer encoder model (Eq. 3).

$$V_{enc} = \text{3D-CNN}(V) \quad (2)$$

$$V_{enc} = \text{TransformerEncoder}(V_{enc}) \quad (3)$$

Frame Encoder. To be able to choose a specific frame as a cover picture, frame-level representations are needed. In our experiments, we sample one of every 25 frames as the cover picture candidates (1 frame per second). We examine the usage of EfficientNet (Tan and Le, 2019) and Vision Transformer (Dosovitskiy et al., 2021) as feature extractors. Both were trained for image classification on ImageNet (Russakovsky et al., 2015). To put the representations into context, we process the sequence of frame features with the Transformer encoder model (Eq. 5).

$$V_{frame} = \text{CNN}(\text{Sample}(V)) \quad (4)$$

$$V_{frame} = \text{TransformerEncoder}(V_{frame}) \quad (5)$$

Before applying the Transformer encoder, we project both the video and frame features into the same dimension as the hidden states of the text encoder. When used in a single model, the two sets of features are concatenated before projecting.

4.3 Interaction Module

Following Yu et al. (2021), who examined different ways of injecting visual information into pre-trained generative language models, we employ the multi-head attention (MHA) based fusion to obtain the vision-guided text representation and perform the fusion after the last encoder layer (Eq. 6–9).

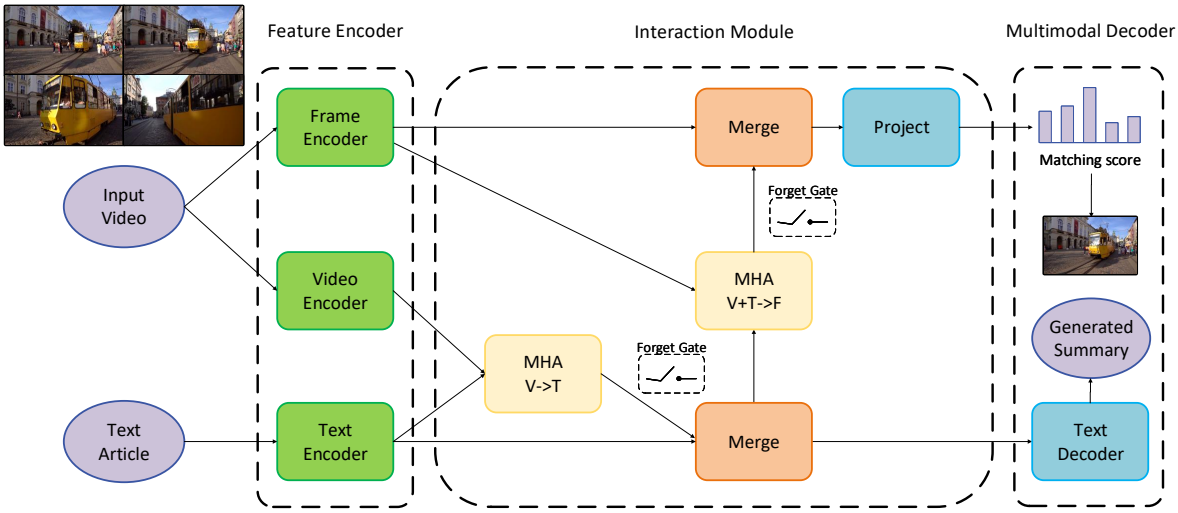


Figure 1: An overview of the proposed MMS model for multimodal summarization.

$$Q = X_{enc}W_q,\ Q \in \mathbb{R}^{M\times d} \quad (6)$$

$$K = V_{enc}W_k,\ K \in \mathbb{R}^{N'\times d} \quad (7)$$

$$V = V_{enc}W_v,\ V \in \mathbb{R}^{N'\times d} \quad (8)$$

$$\widetilde{X}_{enc} = \text{MHA}(Q, K, V),\ \widetilde{X}_{enc} \in \mathbb{R}^{M\times d} \quad (9)$$

As suggested by Liu et al. (2020), we use the forget gate (FG) mechanism so that the model can filter out low-level cross-modal adaptation information (Eq. 10).

$$\widehat{X}_{enc} = \text{FG}(X_{enc}, \widetilde{X}_{enc}),\ \widehat{X}_{enc} \in \mathbb{R}^{M\times d} \quad (10)$$

We use the same MHA mechanism to obtain the text+video guided frame representations $\widehat{V}_{frame}$ by substituting the X_{enc} with V_{frame} in Eq. 6 and V_{enc} with $\widehat{X}_{enc}$ in Eq. 7 and Eq. 8.

4.4 Multimodal Decoder

To generate the textual summary, we use the standard Transformer decoder initializing its weights from the mT5 checkpoint. We use the vision-guided text representation $\widehat{X}_{enc}$ as the input (Eq. 11) and train it using the standard negative log-likelihood loss (NLLLoss) w.r.t. the target sequence Y (Eq. 12).

$$\widehat{Y} = \text{TransformerDecoder}(\widehat{X}_{enc}) \quad (11)$$

$$\mathcal{L}_{text} = \text{NLLLoss}(\widehat{Y}, Y) \quad (12)$$

To obtain the labels C for cover picture (cover frame) selection, we compute the cosine similarity between the CNN features of the reference cover picture and the candidate frames. The similarity of over 99.99% of instances was in the [0,1] range, and the remaining negative values were mapped to 0. The previous works (Li et al., 2020b; Fu et al., 2020) regarded the frame with the maximum cosine similarity as ground-truth and others as negative samples (C_{max}). After examining the cosine similarity patterns (Figure 2), we noticed that the per-video similarity has often either more than one peak, or there are consecutive sequences of frames with very similar scores (capturing a still scene). Our intuition was that this may harm the model performance – very similar frames might be labeled as both positive and negative examples. To overcome this issue, besides the binary labels C_{max}, we introduce the smooth labels C_{smooth} that assign to each frame its cosine similarity score with the reference cover picture.

We use a projection matrix to map the text+video guided frame representations $\widehat{V}_{frame}$ to a single dimension (Eq. 13) and train (Eq. 14) using the binary cross-entropy loss (BCELoss). The target labels C are either C_{max} or C_{smooth}. We train the whole model end-to-end by minimizing the sum of losses $\mathcal{L}$ (Eq. 15).

$$\widehat{C} = \widehat{V}_{frame}W_p,\ W_p \in \mathbb{R}^{d\times 1} \quad (13)$$

$$\mathcal{L}_{image} = \text{BCELoss}(\widehat{C}, C) \quad (14)$$

$$\mathcal{L} = \mathcal{L}_{text} + \mathcal{L}_{image} \quad (15)$$

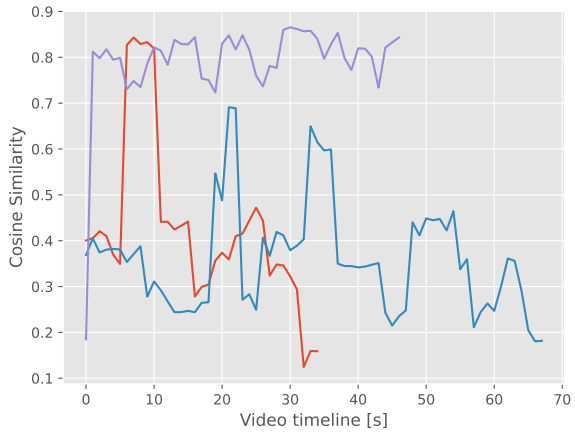


Figure 2: Three examples of cosine similarity plots between CNN features of the reference cover picture and all candidate frames from the video. The examples were chosen manually to present three different video similarity patterns: with a single peak (red), with more than one peak (blue), and with a consecutive sequence of frames having very similar scores (violet).

5 Experiment Setup

5.1 Dataset

In our experiments, we perform the training/dev/test splits of the MLASK dataset following the chronological ordering based on publication date. We use the articles published in the first half (Jan–Jun) of 2021 for validation (2,482 instances) and the ones published in the second half (Jul–Dec) of 2021 and the beginning (Jan–Feb) of 2022 for testing (2,652 instances). The remaining data is used for training (36,109 instances).

5.2 Implementation

We implement our experiments in PyTorch Lightning[3] and use the mT5-small variant (300M trainable parameters) provided via the Transformers (Wolf et al., 2020) package. Following Yu et al. (2021), we use two separate 4-layer encoders with 8 attention heads to contextualize the video and frame representations (Eq. 5 and Eq. 3). As video feature extractors, we use the R(2+1)D 34-layer IG-65M[4] and S3D_HowTo100[5] models to encode sequences of the length of 32 frames. To extract frame-level features, we utilize the EfficientNet-B4 variant from the torchvision package and the vit-base-patch32-224-in21k variant of Vision Transformer provided by Hugging Face (Wolf et al., 2020). We follow the suggested pre-processing (e.g., re-scaling) for each feature extractor independently. The total number of trainable parameters is equal to approximately 323M.

5.3 Hyper-parameters

We train the multimodal model using the Adam optimizer (Kingma and Ba, 2015) with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. We increase the learning rate linearly for the first 8,000 steps (0 to 5e-4) and then follow an inverse square root decay schedule. Since both the text encoder and decoder are pre-trained, we freeze them for the first 2 epochs. We limit the document size to 1,536 sub-word tokens and the summary length to 256 tokens. We train all the models for 50 epochs with an early stopping applied if ROUGE-L (see Section 5.4) does not improve on the dev-set for 5 consecutive epochs. During decoding, we use the best checkpoint with respect to ROUGE-L, utilizing beam search with the beam size of 4, length penalty of 1.0, and repetition penalty (Keskar et al., 2019) of 2.5. We select the cover frame by applying argmax to the projected representations (Eq. 13). We employ gradient accumulation to train with the effective batch size of 32. Each model is trained on a single GeForce RTX 3090 GPU, and the average training time is roughly 36 hours.

5.4 Evaluation Metrics

Most existing implementations of ROUGE (Lin, 2004), a standard metric used to evaluate summarization, are English-specific and utilize e.g., an English stemmer and stop words. Since our dataset is in Czech, following the work of Straka et al. (2018), we evaluate the model performance with language-agnostic variants of ROUGE[6] reporting the F1 scores (ROUGE-1, ROUGE-2, ROUGE-L).

To estimate the quality of cover frame selection, we follow Fu et al. (2020) and report the cosine similarity (CosSim) between the reference cover picture and the chosen cover frame. To have a better understanding of the model performance, we also follow Li et al. (2020b) and report Recall@k (R@k)[7] considering the frame closest to the ground-truth as a positive example. To evaluate the

[3] https://github.com/PyTorchLightning/pytorch-lightning

[4] https://github.com/moabitcoin/ig65m-pytorch

[5] https://github.com/antoine77340/S3D_HowTo100M

[6] https://lindat.cz/repository/xmlui/handle/11234/1-2615

[7] https://github.com/Lightning-AI/metrics

DEV								
	ROUGE-1	ROUGE-2	ROUGE-L	CosSim	R@5	R@10	KC	PC
RandomT	13.92	1.63	9.02	-	-	-	-	-
Lead3	15.47	2.32	10.25	-	-	-	-	-
Oracle	**22.92**	**5.37**	**18.28**	-	-	-	-	-
mT5-MLASK	18.25	4.14	13.07	-	-	-	-	-
mT5-SumeCzech	19.18	4.53	13.76	-	-	-	-	-
RandomV	-	-	-	0.335	0.092	0.182	0.000	0.000
MMS	18.34	4.12	13.26	**0.563**	**0.206**	**0.339**	**0.303**	**0.465**
+ Masked Video	17.70	3.84	12.81	0.548	0.191	0.320	0.275	0.439
− IG-65M	17.74	3.89	12.95	0.558	**0.200**	0.323	0.290	**0.456**
− S3D	17.82	3.88	12.93	0.530	0.187	0.321	0.260	0.428
− Effnet	18.07	4.04	13.13	**0.589**	0.160	0.280	0.211	0.328
− ViT	17.69	3.71	12.82	0.527	0.192	0.320	**0.309**	0.488
+ SumeCzech	19.64	4.95	14.32	0.551	0.192	0.319	0.274	0.440
+ Smooth Labels	**19.73**	**4.97**	**14.34**	**0.562**	**0.202**	**0.332**	**0.295**	**0.458**
+ Masked Video	**19.74**	**5.02**	**14.34**	0.561	0.197	**0.331**	0.290	0.452
TEST								
MMS	18.45	4.29	13.42	**0.552**	0.183	0.321	**0.306**	**0.447**
+ Masked Video	17.65	3.95	12.88	0.542	**0.187**	**0.332**	0.283	0.422
− IG-65M	17.81	4.02	13.07	0.548	**0.186**	0.321	0.296	0.437
− S3D	17.89	4.03	13.03	0.531	0.177	0.316	0.264	0.408
− Effnet	18.21	4.28	13.37	**0.582**	0.157	0.279	0.216	0.311
− ViT	17.78	3.94	13.00	0.509	0.176	0.311	**0.303**	**0.452**
+ SumeCzech	**19.58**	**4.95**	**14.30**	0.541	0.181	0.318	0.278	0.420
+ Smooth Labels	**19.74**	**4.90**	**14.34**	0.551	**0.188**	**0.330**	0.299	**0.444**
+ Masked Video	**19.69**	**4.91**	**14.38**	**0.553**	0.184	**0.326**	**0.300**	0.439

Table 3: Evaluation on the dev-set and test-set of MLASK. See Section 5.4 for the metrics description. The figures are averaged over three runs with different seeds. The three highest-scoring systems in each column are bolded independently for test-set and dev-set.

frame scoring at even coarser video-level granularity, we report Kendall's Tau (KC) and Pearson (PC) correlation coefficients[8] to measure the correlation of ordering based on the projected representations (Eq. 13) with the absolute frame ordering based on similarity with the ground-truth picture.

6 Experiments

We analyze several aspects of the proposed model: First, we study the effect of the visual features. Second, we analyze the contribution of pre-training the model on text-only summarization data. Third, we exploit the smooth frame labels to further improve the model. The results are presented in Table 3.

6.1 Baselines

To put our experiments into context, we first report the performance of several text-only baselines: *RandomT* extracts three random sentences from the article and *Lead3* extracts three initial sentences (trivial baselines); *Oracle* takes three sentences that maximize ROUGE-L with the ground-truth abstract (the upper bound for extractive summarization); *mT5-MLASK* is the output of the mT5 model fine-tuned on the textual part of the MLASK training set and *mT5-SumeCzech* is the mT5 model fine-tuned on the SumeCzech (Section 6.3) dataset (abstractive summarization baselines). There is also a video-only baseline *RandomV*, which performs random frame ordering.

Unsurprisingly, both *mT5* variants outperform the trivial baselines (*RandomT*, *Lead3*), but their results are still far below the *Oracle* performance. Using larger training data (SumeCzech has roughly 20 times more documents than MLASK) improves the performance by approximately 1 ROUGE point.

6.2 Visual features

In Section 4.2, we proposed to employ two different visual features for both video and image feature extraction. The system exploiting all the features is denoted as MMS in Table 3. It achieves slightly higher scores than *mT5-MLASK* (dev-set ROUGE-1: 18.25 → 18.34, ROUGE-L: 13.07 → 13.26) but lags behind the text-only *mT5-SumeCzech* that was trained on a much larger corpus. To analyze the

[8] https://github.com/scipy/scipy

	ROUGE-1	ROUGE-2	ROUGE-L
Lead3	14.34	2.14	9.64
Random3	12.52	1.27	8.37
t2t (2018)	11.30	1.00	8.70
mT5-SumeCzech	**18.46**	**4.54**	**13.33**

Table 4: Performance of the text-to-text summarization models on the test part of SumeCzech (Straka et al., 2018) for the article→abstract task.

effect of the individual visual features, we report the results of the MMS model, excluding those features one by one (see the rows starting with the "−" sign). The scores indicate that the model combining all the features is superior.

6.3 Pre-training

Yu et al. (2021) showed that the usage of pre-trained generative language models is beneficial for multimodal summarization. We explored this idea further by task-specific pre-training on SumeCzech (Straka et al., 2018) – a large-scale Czech news summarization corpus used to fine-tune the mT5 model for summarization (*mT5-SumeCzech*). We used the Adafactor (Shazeer and Stern, 2018) optimizer with a constant learning rate equal to 5e-4 and trained until ROUGE-L ceased to improve on the dev-set for 5 consecutive evaluations. To avoid any training/test data leaks, we excluded from the *mT5-SumeCzech* training data the articles that could appear in MLASK based on the date of publication (794,018 left, i.e., 92%). Performance on the test part of SumeCzech is reported in Table 4. Based on the results (Table 3, system MMS + SumeCzech), we can clearly see that the usage of mT5 fine-tuned for summarization instead of the raw mT5 boosts the performance on the text summarization part (test-set ROUGE-1: 18.45 → 19.58, ROUGE-L: 13.42 → 14.30).

6.4 Smooth labels

In Section 4.4, we proposed to use the smooth labels C_{smooth} during training to overcome the issue of very similar frames being labeled as positive and negative examples. Our results (Table 3, system MMS + SumeCzech + Smooth Labels) indicate that, indeed, this method helps with the quality of cover frame selection (test-set CosSim: 0.541 → 0.551, R@10: 0.318 → 0.330). We can also notice a small improvement (test-set ROUGE-1: 19.58 → 19.74) in the quality of text summarization, which we attribute to more stabilized training.

For a full comparison, we also include two variants with masked video features (MMS + SumeCzech + Smooth Labels + Masked Video and MMS + Masked Video) – all the video features are masked with random noise, both during the training and the evaluation. The frame features are left intact. Surprisingly, for the variant that was pre-trained on a large text-only corpus, masking the video features does not hurt the model performance. This is, however, the case for the model that did not go through the task-specific pre-training. After examining the models, we noticed that the representations after the video encoder (Eq. 5) are not very meaningful, i.e., every segment is mapped to a similar vector. We believe this is due to the indirect usage of video representations in the Cross-modal Interaction Module – too weak learning signal (gradient) is propagated to the video encoder. Considering the drop in performance for the model without pre-training, it seems to be the case that the information from pre-training and multimodal input is not completely orthogonal.

7 Human Evaluation

Previous works on VMSMO evaluated the system performance by employing human judges to assess the quality of generated textual summary: Li et al. (2020b) measured to what extent the system summaries were sufficient to answer questions generated from the reference summary and ranked them based on *Informativeness*, *Coherence*, and *Succinctness*; Fu et al. (2021) scored the system summaries based on *Informativeness* and *Satisfaction*. We believe no prior work employed human annotators to judge the quality of a chosen cover frame (pictorial summary) in the context of textual summary. For the similar task of multimodal summarization with *unimodal output*, Wan and Bansal (2022) collected annotations for the subset of the WikiHow dataset (Yang et al., 2021) that measured whether the textual output was faithful to the source pair of document and image.

7.1 Formulation

To evaluate the quality of cover frame selection, we asked human annotators to judge the quality and usefulness of an image as a pictorial summary of the article. 18 human annotators participated. All were adult, native Czech speakers who read online news magazines daily. Figure 3 displays a screenshot of the annotation tool. For each instance,

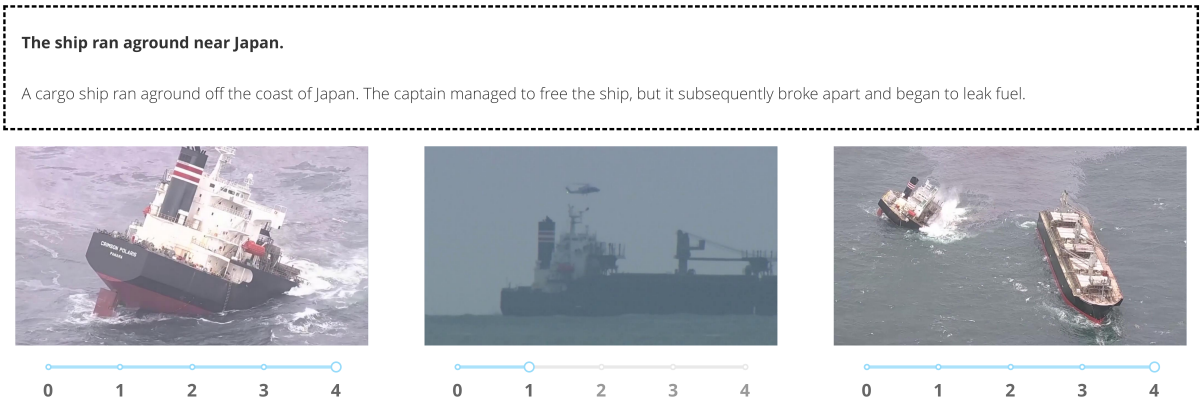


Figure 3: Screenshot of the annotation tool used to collect human judgments about the quality and usefulness of selected cover frame. For convenience, we translated all text into English.

the annotators were asked to rate 3 images on a scale of 0 to 4 (the higher, the better) in the context of the article's title and the reference summary.

The suggested interpretation of the scale levels was:

- 0: The picture is not relevant at all or very marginally (technical quality is not important).
- 1: The image is partly relevant (there is a certain connection between what it captures and the content of the text), but technically imperfect (e.g., blurred, cropped inappropriately, taken from an inappropriate angle or at an inappropriate moment).
- 2: The image is partly relevant (there is a certain connection between what it captures and the text content) and of a good technical quality.
- 3: The picture is very relevant, but technically imperfect (e.g., blurred, cropped inappropriately, taken from an inappropriate angle, or at an inappropriate moment).
- 4: The picture is both very relevant and of a good technical quality. It is a suitable cover picture.

7.2 Setup

We randomly chose 300 instances from the MLASK test-set for annotation and split them into 10 batches of 30 instances. We used the first batch to measure the inter-annotator agreement, asking each annotator to score all the instances in the control batch plus at least in one more.

For each instance, four images were considered for annotation: the reference picture (denoted as *Reference*), a random frame from the video (*RandomV* output), and the outputs of two test models – MMS pre-trained on SumeCzech using the smooth labels (MMS + SumeCzech + Smooth Labels, further denoted as System A) and the same model with masking of the video features (MMS + SumeCzech + Smooth Labels + Masked Video, further denoted as System B). See Appendix A for examples.

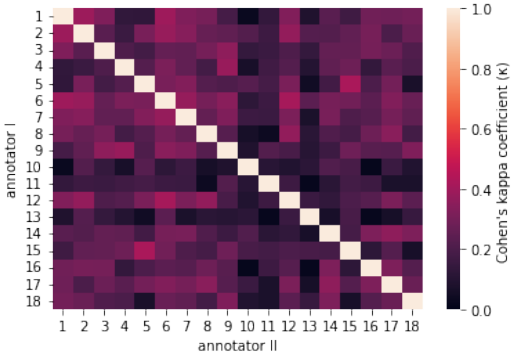


Figure 4: Values of Cohen's κ used to measure the inter-annotator agreement on the control batch (30 instances).

In the control batch, we always included the reference picture, hiding the output from one of the methods in 33% of the cases. In the other batches, we display 3 out of the 4 images selected randomly. To avoid a position bias, we shuffle the images before showing them to the annotator. On average, we collected 2.5 annotations for each image.

7.3 Results

Figure 4 displays the inter-annotator agreement on the control batch in the form of a heat map. The average value of 0.217 indicates a "fair" agreement. One can notice that three annotators (10, 11, and 13) have a lower average agreement (average below 0.2). We decided to exclude their annotations from further analysis. By doing so, the average value of Cohen's κ increased to 0.26, and the average number of annotations decreased to 2.2.

	Total Score	Adequacy Score
Reference	2.89 ± 0.99	1.64 ± 0.50
RandomV	2.39 ± 1.15	1.44 ± 0.61
System A	2.64 ± 1.10	1.51 ± 0.58
System B	2.66 ± 1.04	1.56 ± 0.52

Table 5: System performance on the task of cover picture selection. See Section 7.1 for the label description.

In Table 5, we report the system-level averages of the scores assessed by human annotators (Total Score). On average, the reference picture is assigned the highest score, and our proposed multimodal summarization model performs better than the random baseline. The results of human assessment confirm our previous findings based on automatic metrics – that the model is not utilizing the video features in an effective manner. It is worth noticing, however, that even the reference picture is not considered very relevant (average score below 3) and that none of the differences are statistically significant. To examine the stability of the annotation process, we also report the averages (Adequacy Score) that disregard the quality of the image and focus only on relevance. We do this by mapping the labels from Section 7.1, (i.e., 0 $\rightarrow$ 0; 1 and 2 $\rightarrow$ 1; 3 and 4 $\rightarrow$ 2). The results are in line with the original ones.

8 Conclusions

In this paper, we explored the recently proposed task of video-based multimodal summarization with multimodal output. We extended the available resources to a new language by introducing a multimodal summarization dataset in Czech. We explored the pre-training strategies, showing that transferring knowledge from the simpler, unimodal task of text-to-text summarization helps with the final performance in multimodal settings. We were also able to show that the usage of inner-video similarities, via the introduction of smooth labels during training, helps to stabilize the training. We conducted a human evaluation of the frame-selection process to confirm the quality of the proposed multimodal MMS model. Our findings indicate that the MMS model pre-trained on the text-to-text summarization is not effective in utilizing video features and that future works should carefully examine to what extent the model is able to make use of multimodal input and whether the improvement is orthogonal to e.g., using more data.

Limitations

MLASK dataset collection. While curating the MLASK dataset, we applied a series of rule-based filters (Section 3.1) and collected only those documents that followed a strict HTML structure. No large-scale human evaluation was applied to check the data validity. We sampled a random subset of 100 articles and checked the data preparation and collection manually.

Language and domain bias. We acknowledge that our findings are based on a single dataset, in a particular language (Czech) and from a particular domain (news articles). Due to the novelty of the task, previous datasets proposed for VMSMO (Table 2) are not applicable to our experiments – dataset by Fu et al. (2021) does not provide single cover pictures, and the datasets by Li et al. (2020b) and Tang et al. (2022) are not publicly available. We also acknowledge that due to the data coming from a particular news provider, it may not be free of cognitive biases.

Technical requirements. Considering the modular architecture of the proposed model (Section 4.1), a modern GPU is required for training (using a 24GB GPU we were able to train with a batch size of 2). To store the raw MLASK dataset (videos and images), roughly 750GB of disk space is required.

Human evaluation. While conducting the human evaluation of cover frame selection, we provided a detailed set of instructions (Section 7.1) and used a control batch (30 instances) that was judged by each annotator to compute the inner-annotator agreement. Our findings (Table 5) indicate that there is a certain perception in the data annotation process that we did not analyze – the gold-standard reference picture is, on average, judged as only "partly relevant".

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. LM2018101).

We thank our friends and colleagues that contributed to the annotation process and anonymous reviewers for their valuable feedback.

References

Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. 2021. Video summarization using deep neural networks: A survey. *Proceedings of the IEEE*, 109(11):1838–1863.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Xiyan Fu, Jun Wang, and Zhenglu Yang. 2020. Multi-modal summarization for video-containing documents. *arXiv preprint arXiv:2009.08018*.

Xiyan Fu, Jun Wang, and Zhenglu Yang. 2021. MM-AVS: A full-scale dataset for multi-modal summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5922–5926, Online. Association for Computational Linguistics.

Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. 2019. Large-scale weakly-supervised pre-training for video action recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12038–12047.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Jia-Hong Huang, Luka Murn, Marta Mrak, and Marcel Worring. 2021. Gpt2mvs: Generative pre-trained transformer-2 for multi-modal video summarization. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, ICMR '21, page 580–589, New York, NY, USA. Association for Computing Machinery.

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*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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

Haopeng Li, Qiuhong Ke, Mingming Gong, and Tom Drummond. 2023. Progressive video summarization via multimodal self-supervised learning. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5573–5582.

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 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 4152–4158. AAAI Press.

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.

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.

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.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In *CVPR*.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gülçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shruti Palaskar, Jindřich Libovický, Spandana Gella, and Florian Metze. 2019. Multimodal abstractive summarization for how2 videos. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6587–6596, Florence, Italy. Association for Computational Linguistics.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Milan Straka, Nikita Mediankin, Tom Kocmi, Zdeněk Žabokrtský, Vojtěch Hudeček, and Jan Hajič. 2018. SumeCzech: Large Czech news-based summarization dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.

Peggy Tang, Kun Hu, Lei Zhang, Jiebo Luo, and Zhiyong Wang. 2022. Tldw: Extreme multimodal summarisation of news videos. *arXiv preprint arXiv:2210.08481*.

Daniel Varab and Natalie Schluter. 2021. MassiveSumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10150–10161, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

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.

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.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yue Yang, Artemis Panagopoulou, Qing Lyu, Li Zhang, Mark Yatskar, and Chris Callison-Burch. 2021. Visual goal-step inference using wikiHow. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2167–2179, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. 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.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

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

Below, we attach outputs from the methods that we included in the human evaluation. System A is the MMS + SumeCzech + Smooth Labels model, and System B is the MMS + SumeCzech + Smooth Labels + Masked Video one. For convenience, we include both the original texts in Czech and machine translation to English. The reference text and images were extracted from Novinky.cz.

Reference	*Pětasedmdesátiletý řidič, který při páteční nehodě v Tatcích na Kolínsku usmrtil tříleté dítě a druhé vážně zranil, stejně jako jejich matku, je podezřelý z usmrcení z nedbalosti a z těžkého ublížení na zdraví z nedbalosti. Příčinou nehody byl zřejmě mikrospánek, uvedla v sobotu v tiskové zprávě policejní mluvčí Vlasta Suchánková.* The 75-year-old driver, who killed a three-year-old child and seriously injured the other, as well as their mother, in Friday's accident in Tatky, Cologne, is suspected of negligent homicide and grievous bodily harm. The cause of the accident was apparently microsleep, police spokesperson Vlasta Suchánková said in a press release on Saturday.
System A	*Dva trestné činy usmrcení z nedbalosti vyšetřují od pátečního odpoledne policisté v Nymburce, kde se srazilo osobní auto s dvěma malými holčičkami. Na dotaz Novinek to ve čtvrtek řekla mluvčí policie Eva Suchánková.* Two crimes of culpable homicide have been under investigation since Friday afternoon by police officers in Nymburk, where a car collided with two little girls. Police spokesperson Eva Suchánková told News on Thursday.
System B	*Dvě malé holčičky zemřely v pátek odpoledne při nehodě dvou aut na Nymbursku. Jeden člověk byl vážně zraněn, druhý je ve vážném stavu v nemocnici. Právu to řekla mluvčí policie Eva Suchánková.* Two little girls died in a two-car accident in Nymburk on Friday afternoon. One person was seriously injured, the other is in a serious condition in hospital. Police spokesperson Eva Suchánková told the law.

Table 6: Textual summary – Example 1.

(a) System A

(b) System B

(c) Reference

(d) Random

Figure 5: Pictorial summary – Example 1.

Reference	*Minulý týden v pátek se v Česku začaly oficiálně prodávat nové chytré hodinky Apple Watch Series 7. Zájem byl tak vysoký, že už po pár hodinách zmizely prakticky všechny hodinky z pultů českých obchodů. A vše nasvědčuje tomu, že si případní zájemci budou muset na další várku ještě dlouho počkat.* Last week on Friday, the new Apple Watch Series 7 was officially launched in the Czech Republic. Interest was so high that after a few hours practically all watches disappeared from the counters of Czech shops. And all indications are that potential buyers will have to wait a long time for the next batch.
System A	*V pátek se začaly oficiálně prodávat nové chytré hodinky od společnosti Apple. Zájem o novinku byl tak vysoký, že se v některých případech nepodařilo uspokojit všechny zákazníky, kteří si objednali hodinky s týdenním předstihem.* On Friday, Apple's new smartwatch officially went on sale. Interest in the new smartwatch was so high that in some cases it was not possible to satisfy all customers who ordered the watch a week in advance.
System B	*Zájem o novou generaci chytrých hodinek Watch Series 7 byl tak vysoký, že se v některých případech nepodařilo uspokojit všechny zákazníky, kteří si objednali hodinky s týdenním předstihem. Novinka má být daleko lépe než předchůdce – dostala extrémně tenké rámečky okolo displeje.* Interest in the new generation of Watch Series 7 smartwatches was so high that in some cases it failed to satisfy all customers who ordered a watch a week in advance. The novelty is supposed to be far better than its predecessor – it got extremely thin frames around the display.

Table 7: Textual summary – Example 2.

(a) System A

(b) System B

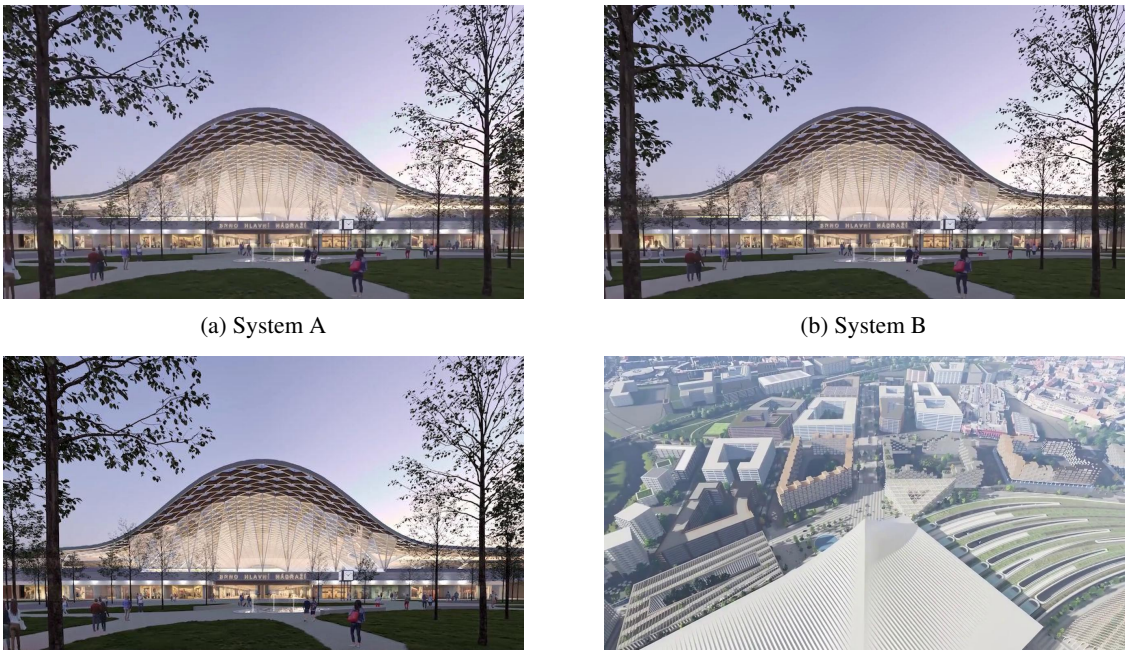
(c) Reference

(d) Random

Figure 6: Pictorial summary – Example 2.

Reference	*Ze století páry rovnou do 21. století. Tak by měla vypadat proměna hlavního brněnského nádraží. Ze současného místa v centru města by se mělo nádraží přesunout o několik set metrů, kde vznikne nový komplex na zelené louce. Jeho podoba by měla být z pera architektů nizozemského ateliéru Benthem Crouwel Architects.* From the century of steam straight into the 21st century. This is what the transformation of Brno's main railway station should look like. From its current location in the city centre, the station should be moved a few hundred metres to create a new green meadow complex. Its design should be from the architects of the Dutch studio Benthem Crouwel Architects.
System A	*Nové hlavní nádraží v Brně by mělo přijít na 50 miliard korun. Vítězný ateliér má zkušenosti z Nizozemska, Nizozemska nebo Nizozemska.* The new main railway station in Brno should cost CZK 50 billion. The winning studio has experience from the Netherlands, the Netherlands or the Netherlands.
System B	*Nové hlavní nádraží v Brně by mělo být hotové do roku 2035. Vítězný ateliér, který má s nádražím bohatou zkušenost, dostal od města peníze za 50 miliard korun.* The new main railway station in Brno should be ready by 2035. The winning studio, which has extensive experience with the station, received money from the city worth CZK 50 billion.

Table 8: Textual summary – Example 3.



(a) System A

(b) System B

(c) Reference

(d) Random

Figure 7: Pictorial summary – Example 3.