

# A picture is worth a thousand words - on Multimodal Summarization

Mateusz Krubiński



Charles University  
Faculty of Mathematics and Physics  
Institute of Formal and Applied Linguistics



unless otherwise stated


# Agenda


- 1. Introduction to Multimodal Summarization**
  - a. What is the task about?**
- 2. Multimodal Summarization in the wild**
  - a. How exactly is the task framed?**
  - b. How did it evolve over the time?**
- 3. Future Directions**
  - a. What are the biggest challenges and open problems?**

# Introduction

# Multimodal? Summarization?

## modality *noun*

 /məʊ'dæləti/

 /məʊ'dæləti/

(plural **modalities**)


- 1 ★ [countable] (*formal*) the particular way in which something exists, is experienced or is done
  - *They are researching a different modality of treatment for the disease.*
- 2 ★ [uncountable] (*linguistics*) the idea expressed by modals
- 3 ★ [countable] (*biology*) the kind of senses that the body uses to experience things
  - *the visual and auditory modalities*

## summary *noun*



OPAL W

OPAL S

 /'sʌməri/

 /'sʌməri/

(plural **summaries**)

- ★ a short statement that gives only the main points of something, not the details

## modalność

Wielki słownik ortograficzny PWN\*

modal•ność -ści

Słownik języka polskiego PWN\*

### modalny

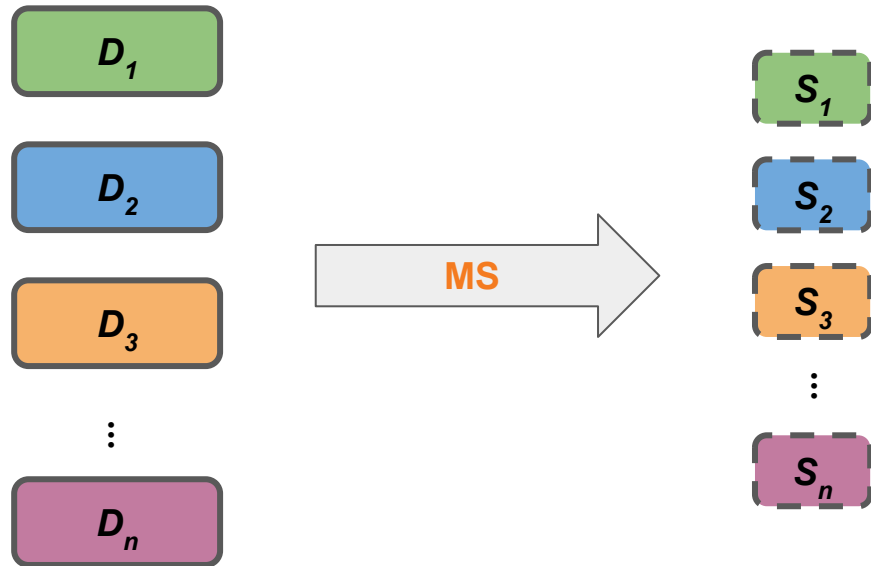
1. «w ontologii: dotyczący sposobu istnienia rzeczy lub zjawisk»
  2. «w logice: odnoszący się do stopnia pewności sądu»
  3. «dotyczący stosunku osoby mówiącej do treści jej wypowiedzi, wyrażanego za pomocą środków językowych»
  4. «mający związek z archaicznymi skalami kościelnymi»
  5. «w statystyce: wartość cechy występującej w danej grupie najczęściej»
- modalność

## streszczenie

Słownik języka polskiego PWN\*

**streszczenie** «treść czegoś ujęta krótko, zwięźle»

# Multimodal Summarization - formulation



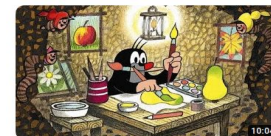
Audio, Video, Text, ...

- $D_i \neq D_j$  and  $D_i \approx S_i$ 
  - Our information is multimodal and no new modalities are generated
- $|D_i| > |S_i|$ 
  - We want to “compress” the data
- $M(D_i) \sim M(S_i)$ 
  - ... so that it is still “meaningful”
- $S_i \stackrel{?}{\neq} \emptyset$ 
  - Not all input modalities must be present in the output
  - A single  $\{S_i\}$  - *unimodal output*
  - More than one  $\{S_i, \dots, S_j\}$  - *multimodal output*

# Multimodal Summarization - single modality



Video → Image



**Abstract** While many approaches to make neural networks more fathomable have been proposed, they are restricted to interrogating the network with input data. [...] In this work, we propose neural persistence, a **complexity measure for neural network** architectures based on **topological** data analysis on weighted stratified graphs. [...]

**Intro** [...] In this work, we present the following contributions: We introduce neural persistence, a novel measure for characterizing the structural **complexity of neural networks** that can be efficiently computed. [...]

**Conclusion** [...] However, this did not yield an early stopping measure because it was never triggered, thereby suggesting that neural persistence **captures salient information** that would otherwise be hidden among all the weights of a network [...]

Text → Text

**TLDR** We develop a new **topological complexity measure for deep neural networks** and demonstrate that it **captures their salient properties**.

# Multimodal Summarization - multimodal output

## Utknął po szyję w bagnie. Odwdzięczył się za ratunek [WIDEO]

Niesamowita akcja ratunkowa na Pomorzu. Myśliwi i strażacy potrzebowali dwóch godzin, aby uwolnić łosia, który po szyję ugrzązł w bagnie w okolicach Człuchowa. Po zakończonej sukcesem operacji zwierzę elegancko podziękowało swym wybawcom.



Łoś bez pomocy nie przeżyłby (Screen: Facebook/ZD PZL Słupsk)

"Członkowie kół łowieckiego »Szarak« Człuchów uratowali łoszcza, który ugrzązł w bagnie. Cała akcja ratunkowa trwała ponad 2 godziny i zakończyła się happy endem" – opisują w mediach społecznościowych całe zajście myśliwi.

## "Musielśmy użyć pasów transportowych"

Na łosia, który utknął w błocie, natknął się członek kół Paweł Obarzanek. Potężne zwierzę ppo szyję zatopił się w błocie. Dlatego myśliwiy wraz z dwoma kolegami ruszyli na pomoc. Gdy zdali sobie sprawę, że sami nie uratują zwierzęcia, wezwali pomoc.

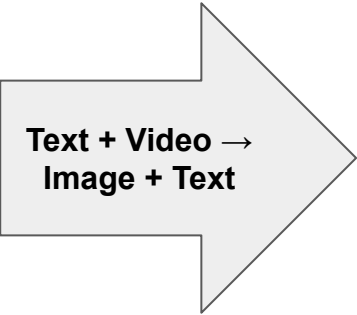


— Na miejsce wysłaliśmy zastęp strażaków, który zastał trzech członków kół łowieckiego, którzy wezwali pomoc. Nasze działania sprowadzały się do wybrania części przemokniętej ziemi i wyciągnięcia łosia za pomocą pasów transportowych na brzeg mokradła — relacjonuje w rozmowie z Onetem mł. brzyg, Piotr Frankenstein z Komendy Powiatowej PSP w Człuchowie.

• Rosyjski samolot "znalazł" 12 nieżyjących boeingów i airbusów w swoich hangarach

## "Łoś podziękował wzrokiem"

Jak mówi w rozmowie z radiem Weekend FM łowczy Kół Łowieckiego "Szarak" w Człuchowie Andrzej Bachan, po wydostaniu się łos kilkakrotnie próbował odejść, ale musiał odpocząć. Jednak po kilku minutach zwierzę powstało, wymownie spojrzelo w stronę ratowników i odeszło do lasu. Całą akcję z oddalenia miała obserwować dorosła samica łosia z młodym.



Łoś utknął po szyję w bagnie. Tak odwdzięczył się za ratunek [WIDEO]

# Multimodal Summarization - unimodal output



Audio + Text → Text

## Call to order

*Facilitated by the chair of the board.*

[Chair of the board's name] called to order the regular meeting of [your organization] at [time of meeting] on [date of meeting] in [location of meeting].

## Attendance

*Facilitated by the secretary.*

Secretary conducted roll call. The following persons were present:

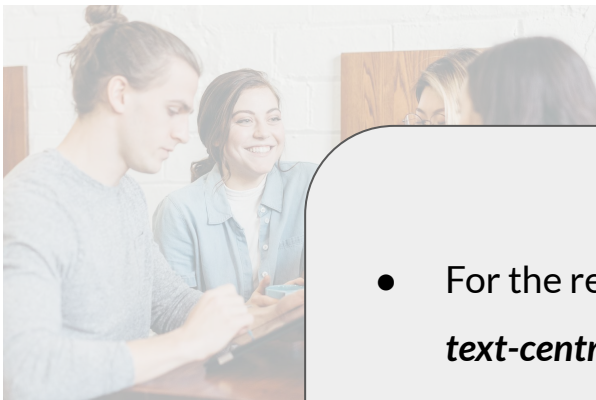
- [Name]
- [Name]
- [Name]

The following persons were absent:

- [Name]
- [Name]
- [Name]



# Multimodal Summarization - unimodal output



- For the remainder of this talk, we will focus only on **text-centric Multimodal Summarization**
- We require that the **text modality is present both in the input and in the output**

Call to order

*Facilitated by the chair of the board*

The regular meeting of [your  
meeting] in [location of meeting].

Persons were present:

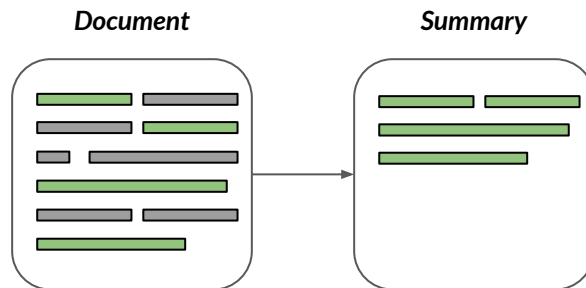
• [Name]

# Formulation and history

# Text Summarization - brief recap

## Abstractive:

- Supervised:
  - Sequence-to-sequence models
- Unsupervised:
  - Foundation models



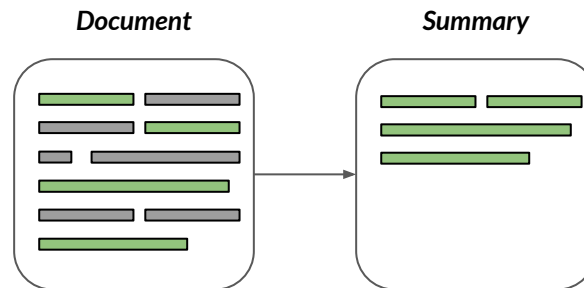
## Extractive:

- Supervised:
  - Sentence classification
- Unsupervised:
  - Graph-based methods

# Text Summarization - brief recap

## Abstractive:

- Supervised:
  - Sequence-to-sequence models
- Unsupervised:
  - Foundation models



## Extractive:

- Supervised:
  - Sentence classification
- Unsupervised:
  - Graph-based methods

## Domains:

- News Articles
- Scientific Papers
- Dialogues
- Meetings
- Multiple Documents
- Medial Reports
- ...

## EVALUATION

### Surface-level:

- ROUGE- $\{1,2,L,S\}$  (Lin, 2004)
- BLEU (Papineni, 2002)
- CHRF (Popovič, 2015)

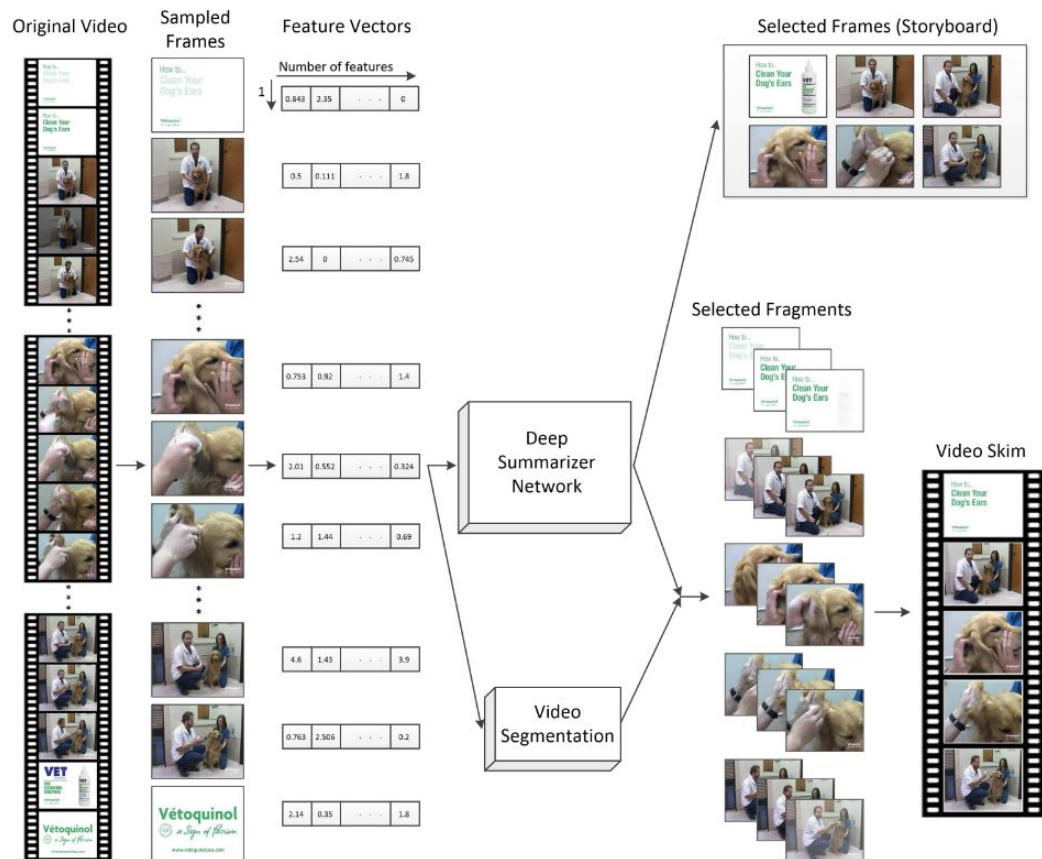
### Model-based:

- MoverScore (Zhao, 2019)
- BERTScore (Zhang, 2019)
- InfoLM (Colombo, 2022)
- COMES (Krubinski, 2022)

### Question Answering-based:

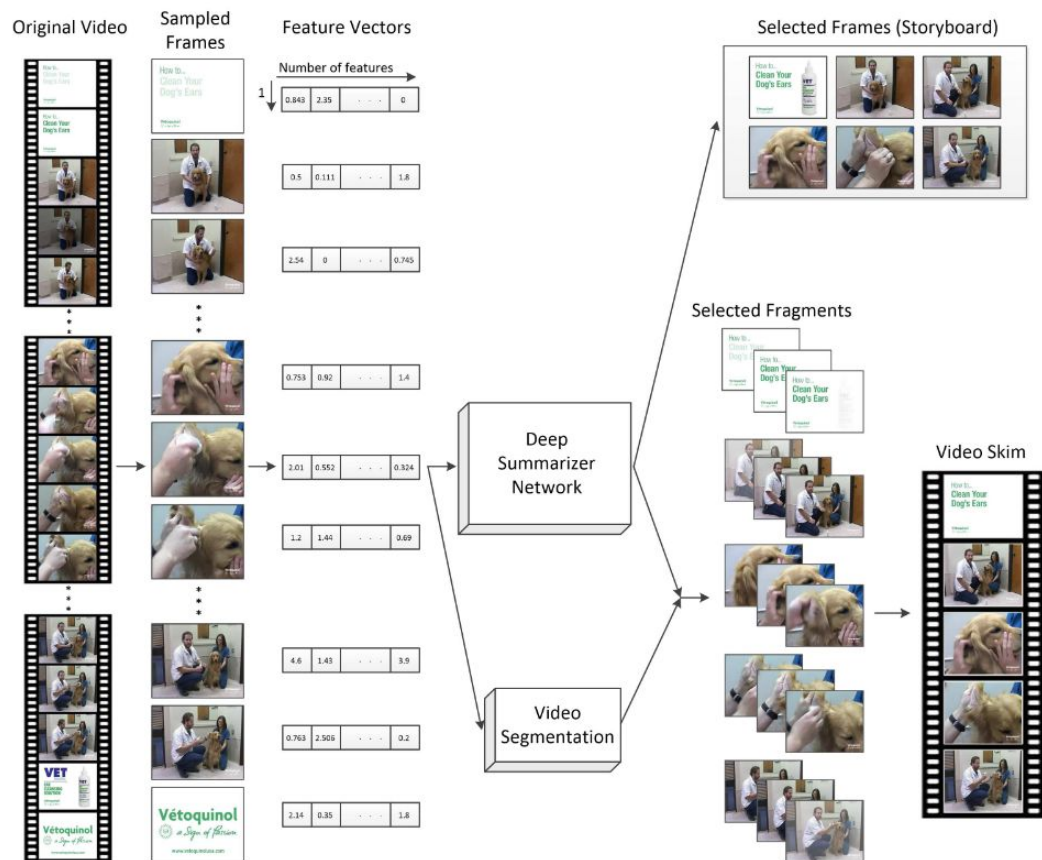
- SummaQA (Scialom, 2019)
- APES (Eyal, 2019)
- QAGS (Wang, 2020)
- QuestEval (Scialom, 2021)

# Video Summarization - brief recap



- Extractive approach
- Generic pipeline of:
  - Down-sample frames
  - Extract frame-level features
  - Model intra-video frame importance
  - Supervised - frame-level annotations, unsupervised - GAN/VAE

# Video Summarization - brief recap

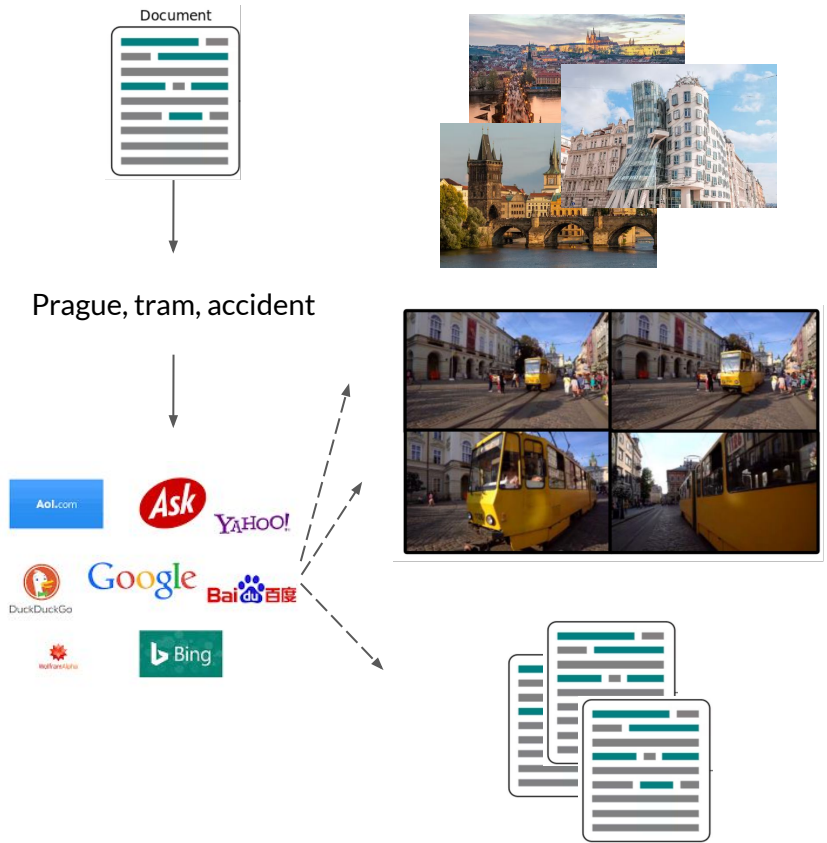


- Extractive approach
- Generic pipeline of:
  - Down-sample frames
  - Extract frame-level features
  - Model intra-video frame importance
  - Supervised - frame-level annotations, unsupervised - GAN/VAE

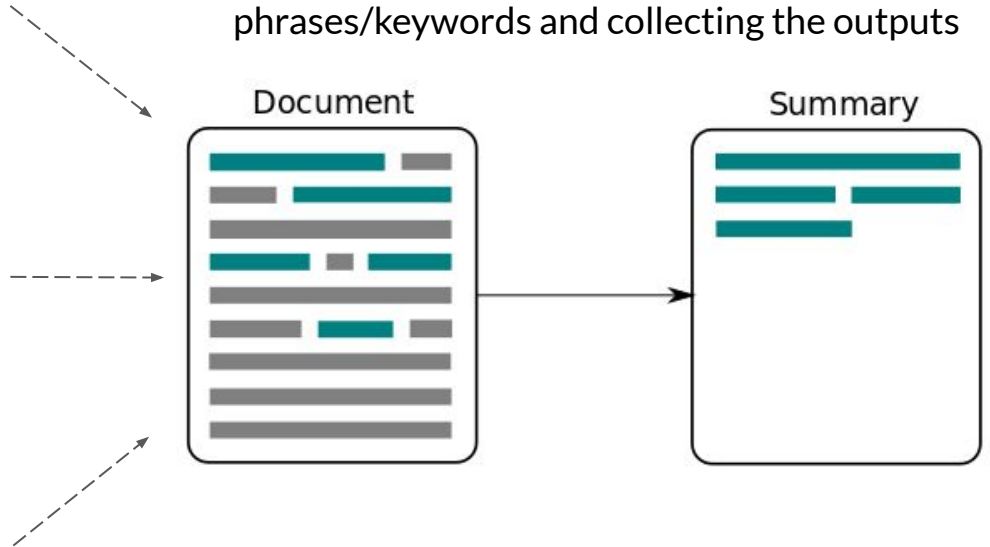
Dataset	# of instances	Modality
TVSum	50	Video
SumMe	50	Video
OVP	50	Video
LoL	218	Video
CNN/DailyMail	312,000	Text
XSum	204,000	Text
Gigaword	3.8M	Text

# Extending Textual Data

Li et al. (2017),  
Li et al. (2018)



- **Original data – text only**
- Additional, multimodal data is obtained by querying a search engine with related phrases/keywords and collecting the outputs



<https://hackernoon.com/summarization-with-wine-reviews-using-spacy-b49f18399577>  
<https://wildstonesolution.com/top-search-engines/>  
<https://unsplash.com>

# Multimodal Documents

**Transcript**

today we are going to show you how to make spanish omelet . i 'm going to dice a little bit of peppers here . i 'm not going to use a lot , i 'm going to use very very little . a little bit more then this maybe . you can use red peppers if you like to get a little bit color in your omelet . some people do and some people do n't .... t is the way they make there spanish omelets that is what she says . i loved it , it actually tasted really good . you are going to take the onion also and dice it really small . you do n't want big chunks of onion in there cause it is just pops out of the omelet . so we are going to dice the up also very very small . so we have small pieces of onions and peppers ready to go .

**Video**



**Summary**

how to cut peppers to make a spanish omelette; get expert tips and advice on making cuban breakfast recipes in this free cooking video .

**Original data is multimodal.**

- News articles
- Instructional video
- Meeting recordings
- Sport events
- Product reviews
- ...

Sanabria et al. (2018), Palaskar et al. (2019), Li et al. (2020), Im et al. (2021)

Field name	Field value
Name	The Hush Puppy
Categories	Southern, Seafood, Restaurants
Noise level	average
Alcohol	beer and wine
⋮	⋮
Ratings	4



Review 1	The fresh water catfish is probably the best I've ever had. The service was outstanding. I would recommend this little secret to everyone.
Review 2	I loved everything about this place!! Great food, great decor, and great service. The best collard greens I have ever had. We had fried oysters for a starter and although I have never had them before so I have nothing to compare them with they were very tasty. The warm hush puppies with the honey butter was delicious!! I had the crab legs which were perfect and plentiful. My sister had the all you can eat fried catfish that was also cooked perfectly. A great experience all around!!
Review 3	Amazing food and great service! The hospitality was out of this world. Will definitely be back soon. The wait was less than 5 minutes at 7pm on a Friday night, amazing!! The staff was very kind and the waitresses were very attentive and helpful. We tried the frog legs, catfish, alligator bites, crab legs, gumbo and of course the hush puppies! Everything was outstanding. What a hidden gem!
Review 4	I love this place the food amazing the staff helpful ....must try green tomatos ...fresh water fish :')
Review 5	We love this place the catfish is good the hush puppies with that honey butter are awesome the french fries the gumbo what else is good there the alligator tail mostly everything on the menu. I guess the only bad thing I can say is sometimes it's like a 20 minute wait in the drive-through but it's well worth it when your food is hot Because tonight I got to go home and warm it up it's not hot enough. Even though they're still open for another hour that was a bummer
Review 6	Really tasty catfish, shrimp and fixin's. Our friend took us to the sister location on Nellis a couple of months ago, but this location was more convenient to our hotel. No worries, this place was just as good! Excellent service, and the salad bar is a nice touch as well. As a Bostonian, I'm pretty particular about seafood. The Hush Puppy fits the bill. Very satisfied!
Review 7	First Time here and the food, staff was awesome. Manager came over and gave us samples of the fried catfish, super nice.
Review 8	I never eat catfish. It's nasty to me until I tasted the saltwater catfish!!! Greens are on point. The hushpuppy are bomb with honey butter!!!! Gator bites where are ok.
Copycat	This place is awesome! The food was great, the service was great. We had the catfish po'boy and it was delicious. The only reason I didn't give 5 stars is because of the fact that they don't deliver.
Self & Control	I love this place. The service is awesome. The hush puppies are to die for. I love the honey butter. I can't wait to go back and try it again. The only thing I don't like about the place is the wait. It can be a little long, but it's worth it. It's a little on the pricey side, but you're getting what you pay for. Love the hot butter, the hush puppies, the French fries, the gumbo, the catfish and the gumbo. Everything is so yummy and the service is top notch. Try it out, you won't be disappointed.
MultimodalSum	This place is a hidden gem. The food is great and the service is even better. I had the all you can eat catfish and it was delicious. The hush puppies are the best I've ever had. I will definitely be back.
Gold	Yummy and delicious catfish. You gotta try it. Friendly staff and service is good too. You can tell they know their seafood and how to prepare and cook it to perfection. The staff also answered any questions I had. The Hush Puppies are tasty too.



# Multimodal Documents

**Transcript**

today we are going to show you how to make spanish omelet . i 'm going to dice a little bit of peppers here . i 'm not going to use a lot , i 'm going to use very very little . a little bit more then this maybe . you can use red peppers if you like to get a little bit color in your omelet . some people do and some people do n't .... t is the way they make there spanish omelets that is what she says . i loved it , it actually tasted really good . you are going to take the onion also and dice it really small . you do n't want big chunks of onion in there cause it is just pops out of the omelet . so we are going to dice the up also very very small . so we have small pieces of onions and peppers ready to go .

**Video**



**Summary**

how to cut peppers to make a spanish omelette; get expert tips and advice on making cuban breakfast recipes in this free cooking video .

**Original data is multimodal.**

- News articles
- Instructional video
- Meeting recordings
- Sport events
- Product reviews
- ...

Sanabria et al. (2018), Palaskar et al. (2019), Li et al. (2020), Im et al. (2021)

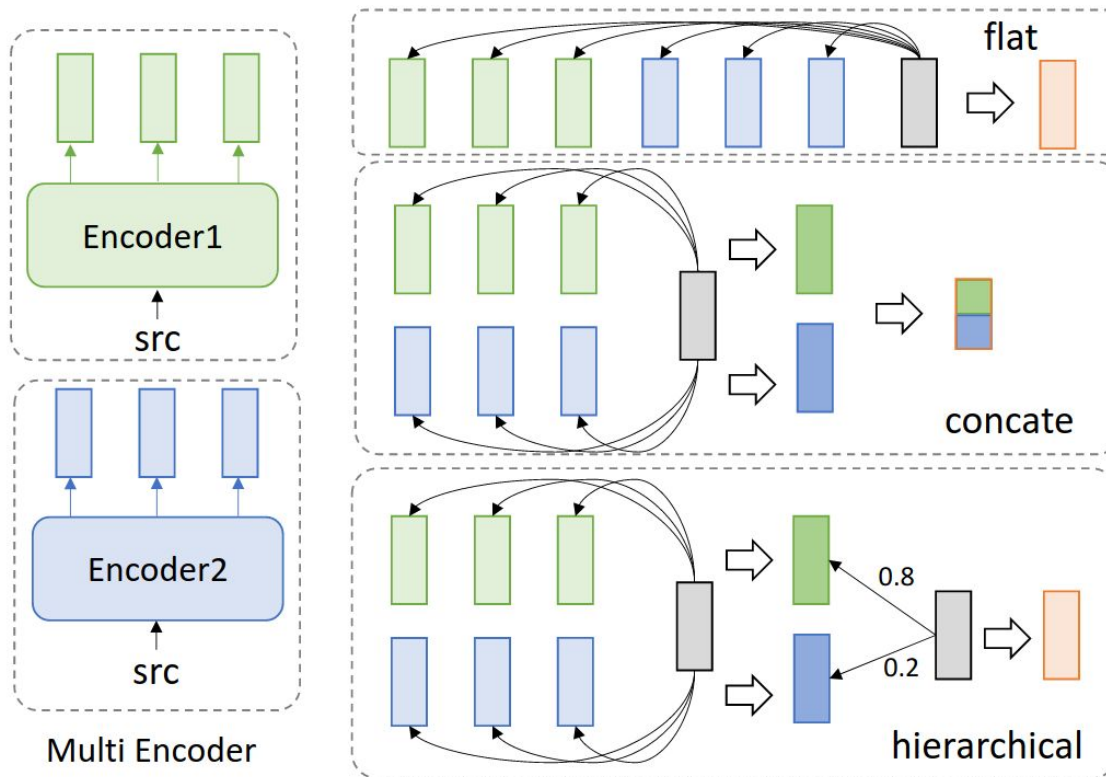
Field name	Field value
Name	The Hush Puppy
Categories	Southern, Seafood, Restaurants
Noise level	average
Alcohol	beer and wine
⋮	⋮
Ratings	4



Review 1	The fresh water catfish is probably the best I've ever had. The service was outstanding. I would recommend this little secret to everyone.
Review 2	I loved everything about this place!! Great food, great decor, and great service. The best collard greens I have ever had. We had fried oysters for a starter and although I have never had them before so I have nothing to compare them with they were very tasty. The warm hush puppies with the honey butter was delicious!! I had the crab legs which were perfect and plentiful. My sister had the all you can eat fried catfish that was also cooked perfectly. A great experience all around!!
Review 3	Amazing food and great service! The hospitality was out of this world. Will definitely be back soon. The wait was less than 5 minutes at 7pm on a Friday night, amazing!! The staff was very kind and the waitresses were very attentive and helpful. We tried the frog legs, catfish, alligator bites, crab legs, gumbo and of course the hush puppies! Everything was outstanding. What a hidden gem!
Review 4	I love this place the food amazing the staff helpful ....must try green tomatos ...fresh water fish :)
Review 5	We love this place the catfish is good the hush puppies with that honey butter are awesome the french fries the gumbo what else is good there the alligator tail mostly everything on the menu. I guess the only bad thing I can say is sometimes it's like a 20 minute wait in the drive-through but it's well worth it when your food is hot Because tonight I got to go home and warm it up it's not hot enough. Even though they're still open for another hour that was a bummer
Review 6	Really tasty catfish, shrimp and fixin's. Our friend took us to the sister location on Nellis a couple of months ago, but this location was more convenient to our hotel. No worries, this place was just as good! Excellent service, and the salad bar is a nice touch as well. As a Bostonian, I'm pretty particular about seafood. The Hush Puppy fits the bill. Very satisfied!
Review 7	First Time here and the food, staff was awesome. Manager came over and gave us samples of the fried catfish, super nice.
Review 8	I never eat catfish. It's nasty to me until I tasted the saltwater catfish!!! Greens are on point. The hushpuppy are bomb with honey butter!!!! Gator bites where are ok.
Copycat	This place is awesome! The food was great, the service was great. We had the catfish po' boy and it was delicious. The only reason I didn't give 5 stars is because of the fact that they don't deliver.
Self & Control	I love this place. The service is awesome. The hush puppies are to die for. I love the honey butter. I can't wait to go back and try it again. The only thing I don't like about the place is the wait. It can be a little long, but it's worth it. It's a little on the pricey side, but you're getting what you pay for. Love the hot butter, the hush puppies, the French fries, the gumbo, the catfish and the gumbo. Everything is so yummy and the service is top notch. Try it out, you won't be disappointed.
MultimodalSum	This place is a hidden gem. The food is great and the service is top notch. Had the all you can eat catfish and it was delicious. The hush puppies are the best I've ever had. I will definitely be back.
Gold	Yummy and delicious catfish. You gotta try it. Friendly staff and service is good too. You can tell they know their seafood and how to prepare and cook it to perfection. The staff also answered any questions I had. The Hush Puppies are tasty too.

# Multimodal Documents

Sanabria et al. (2018),  
Palaskar et al. (2019),  
Li et al. (2020),  
Im et al. (2021)



## Modeling techniques

- Extending seq-2-seq with attention over the visual input
- Realised with separate encoders
- Representations are merged before decoder

# How to summarize multimodal documents?

## REVEALED: NASA's full picture set from James Webb Telescope will show detailed views of stellar nurseries with stars larger than the sun and a galaxy group 290 million light-years away

- NASA's James Webb Telescope will show new views of stellar nurseries, a galaxy group and a huge planet outside our solar system
- The space agency lists five targets for the first set of full-color scientific images being released on Tuesday, July 12 at 10:30 am EDT
- 'I'm as excited as everyone else who is anticipating the release of the first beautiful full-color images and data,' said a longtime Webb scientist
- The release of the first images is just the beginning of Webb's scientific operations as it seeks to 'unfold the universe'

PUBLISHED: 18:24 BST, 8 July 2022 | UPDATED: 18:27 BST, 8 July 2022

NASA revealed the James Webb Telescope will target multiple spectacular cosmic objects - including far-flung stellar nurseries, a giant planet outside of our solar system and a galaxy group that's 290-million light-years away - ahead of the release of its first images.

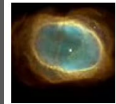
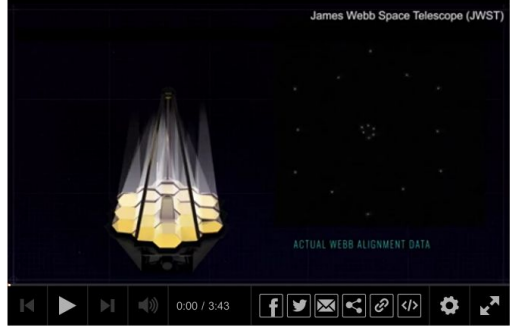
The space agency lists five main targets for the \$10 billion telescope's first set of full-color scientific images being released on Tuesday, July 12 at 10:30 am EDT.

'Even after working on the program for many years, I'm as excited as everyone else who is anticipating the release of the first beautiful full-color images and data from NASA's James Webb Space Telescope - an audacious endeavor in partnership with the European and Canadian space agencies,' says Eric Smith, a Webb program scientist at NASA who has been working on the telescope team since its beginnings in the mid-1990s.



© NASA  
'The James Webb Space Telescope will give us a fresh and powerful set of eyes to examine our universe,' Webb program scientist Eric Smith said. Pictured is the Carina Nebula as seen from NASA's Hubble Space Telescope

### NASA reveals James Webb's first focused image of a single star



Science

Article

## REVEALED: NASA's full picture set from James Webb Telescope will show detailed views of stellar nurseries with stars larger than the sun and a galaxy group 290 million light-years away

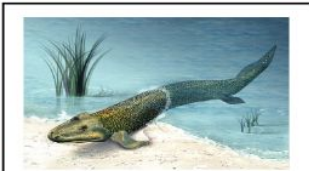
By Christopher Carbone For Dailymail.Com - July 8th 2022, 6:24:27 pm

NASA's James Webb Telescope will target multiple cosmic objects - including stellar nurseries, a giant planet and a galaxy group that's 290-million light-years away.

# Multimodal Summarization with Multimodal Output



Researchers have discovered the fossilized remains of a small, lizard-like creature that is the missing ancestral link ...



Tiny was one of the first **four-legged creatures** to move ...

Zhu et al. (2018),  
Li et al. (2020),  
Fu et al. (2021),  
Tang et al. (2022)



**News:**  
A software company in New Zealand began training robots to graze sheep. The current training results are performing well, and the sheep can accept most commands issued by the robot.



Shepherds are about to lose their jobs

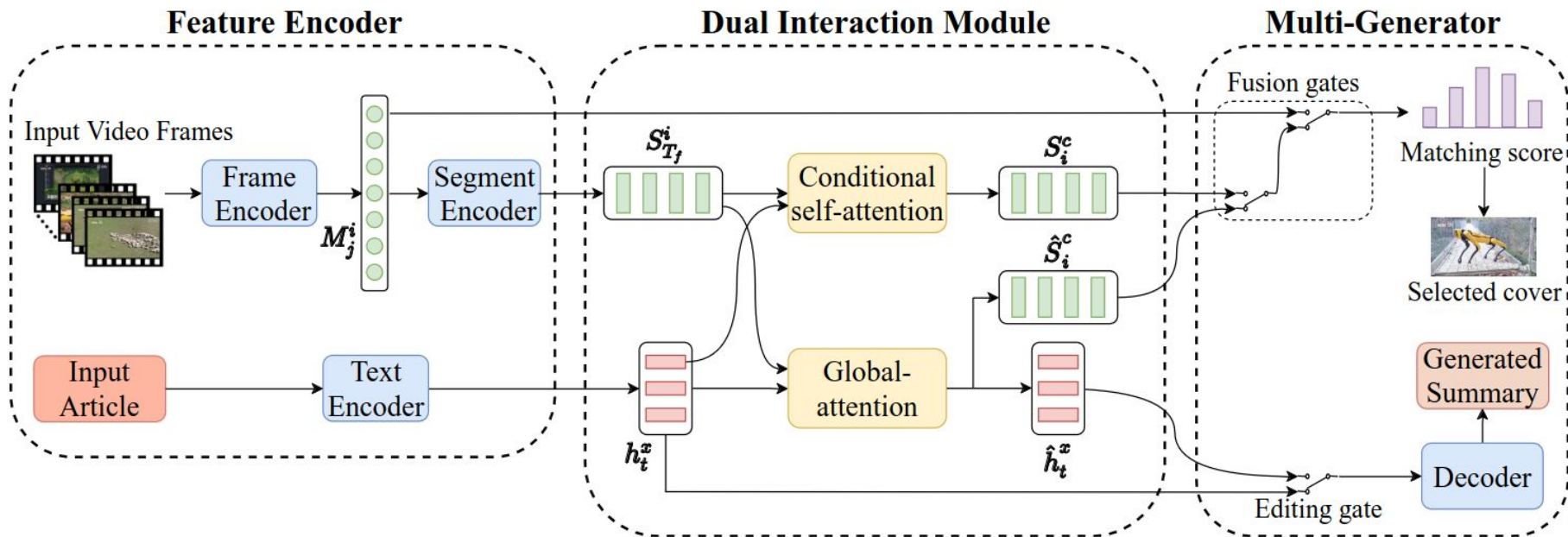
**Article:** Community members in Brentwood, New York on Long Island are outraged that police have not yet publicly named a suspect in the death of Evelyn Rodriguez, 50. Rodriguez was struck by a vehicle and killed on September 14 in cul-de-sac near where her daughter's dead body was found, two years ago to the day. Her daughter, Kayla Cuevas, 16, and her best friend, Nisa Mickens, 15, were found in the wooded area behind one of the homes in the circle in 2016. ... The brutal murders of Nisa Mickens and Kayla Cuevas, and the savage killing of Jose Pena, allegedly committed by these defendants, exemplify the depravity of a gang whose primary mission is murder,' Capers said. (about 757 tokens)



**Reference:** Evelyn Rodriguez, 50, was run down and killed on September 14, while preparing for a candlelight vigil for her murdered daughter, Kayla Cuevas, 16 Cuevas and her best friend, Nisa, 15, were found beaten and hacked to death, allegedly by members of the notoriously violent MS-13 gang, in 2016.

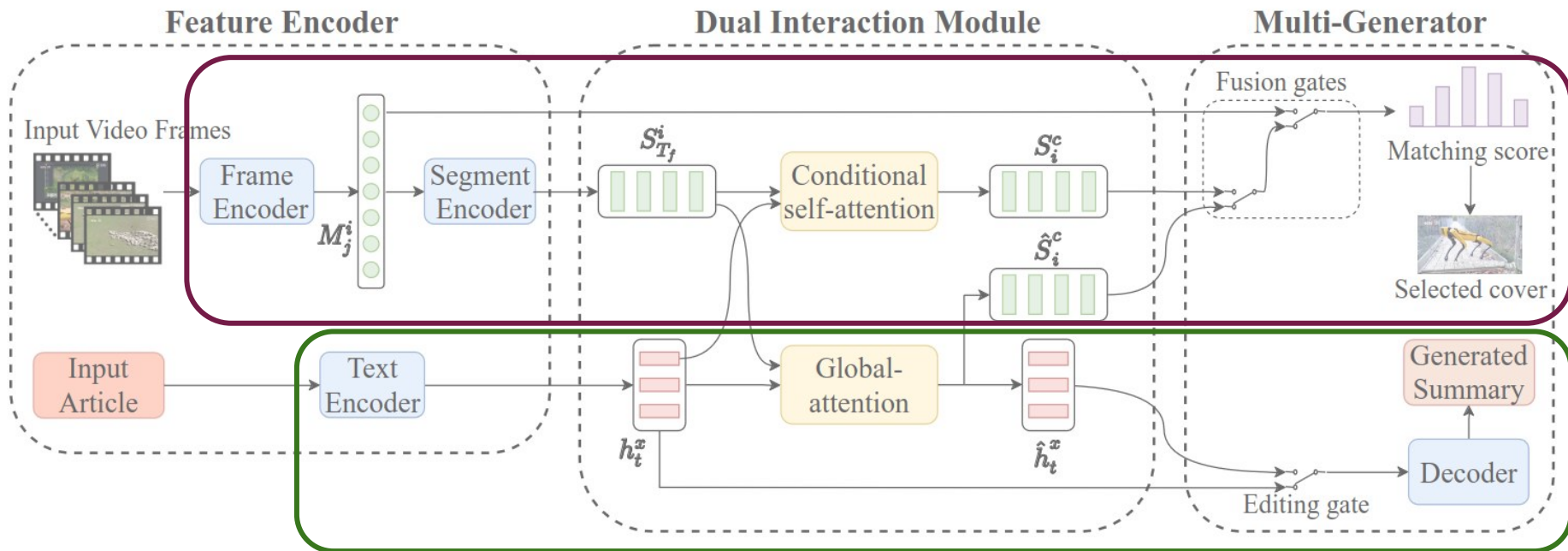


# Multimodal Summarization with Multimodal Output



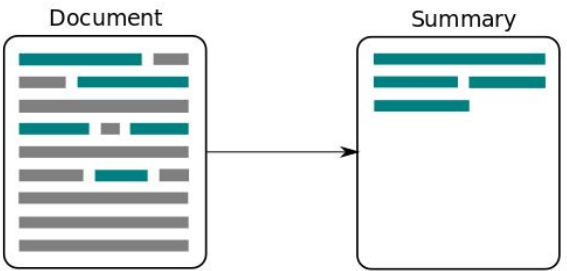
# Multimodal Summarization with Multimodal Output

- Text-aware frame scoring
- Dual encoder, merged before decoder



# Multimodal Summarization - task evolution

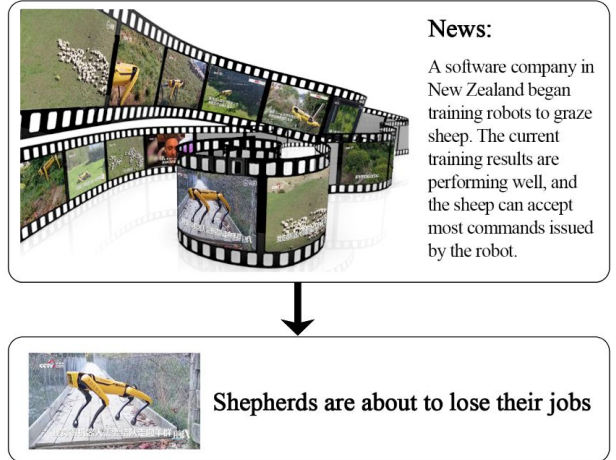
## 1) Text -> Text



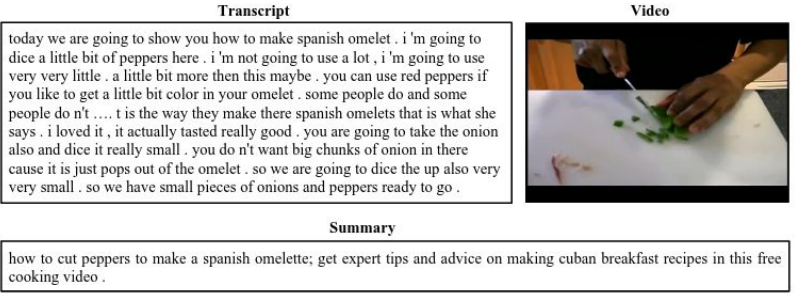
## 2) Text + (external) V/I/T -> Text



## 4) Text + V/I -> Text + I



## 3) Text + V/I -> Text



## Challenges and open problems



# Challenges and open problems

1. **Lack of data**
2. **Multimodal feature extraction and cross-modal fusion**
3. **Formulation of learning signal and task-specific pre-training**
4. **Multimodal evaluation**

## **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



**EACL 2023**

# Lack of data

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
<b>MLASK (this paper)</b>	41,243	277	33	86s	Czech

- Video ■
- Article ■
- Title ■
- Summary ■
- Image Summary ■



Vláda strop na energie zaozkrouhila. Bude o něco vyšší



Nahrávka: pozor na podvodníky, feld Babiš. Rozzúřil ho znalec, který podpořil obžalobu



Je to podvodník, řekl Babiš. Rozzúřil ho znalec, který podpořil obžalobu



Sklarešly vyhlížeji strop na energii. Bez něj hrozí nejdražší sezona




Šéf Siko: Máme na přítel rok tři scénáře, ale ani jeden není optimistický



Atlas vlivu mapuje, jak politici šíří ruský a čínský vliv. Vedou SPD a ANO

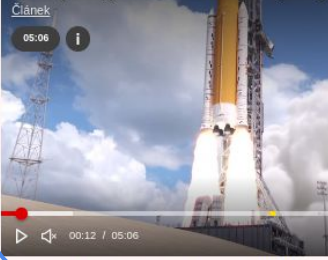
Zprávy » Tech » Technologie » NASA podruhé zrušila start rakety k měsíci. Podívejte se, jak měl...

## NASA podruhé zrušila start rakety k měsíci. Podívejte se, jak měla letět

 JAN MAREK f t

Podívejte se, jak má k Měsíci letět nejsilnější a nejdražší raketa n...

Článek



Nejsilnější raketa Země dnes letí k Měsíci

BRÁNO NOSTR: SLS (Space Launch System)

VÝŠKA: 98,1 m

KAPACITA PŘI NÁKLAD K MĚSÍCI: 26,9 t

05:06

00:12 / 05:06


Podívejte se, jak má k Měsíci letět nejsilnější a nejdražší raketa na světě. | Video Jan Marek | 3. 9. 17:54


Americký Národní úřad pro letectví a vesmír (NASA) dnes podruhé v tomto týdnu kvůli technickým potížím zrušil plánovaný start rakety SLS s modulem Orion k Měsíci. Vesmírná agentura o tom informuje na svých internetových stránkách.


Dnes ve 20:17 SELČ se mělo na dvě hodiny otevřít okno pro start rakety SLS s modulem Orion k Měsíci. NASA už předtím na svých stránkách uvedla, že má potíže s únikem paliva a o několik hodin později start zcela zrušila.

Šlo během jednoho týdne o už druhý pokus o vzlet vesmírného korábu s největším tahem, ale i cenou na světě. Jeho vývoj se totiž dost prodloužil i kvůli problému s motory, které přerušily start k Měsíci i toto pondělí. Brzy má přitom letět i s lidmi.

### STALO SE

VCERA 22:31  
**Investoři se lekli inflace. Americké akcie klesly nejvíce za dva roky** 

VCERA 21:56  
**Královská rodina se rozloučí s královnou. Rakev dorazila do Buckinghamu** 

VCERA 20:56  
**Video: Drony na vodík z výstavy novinek jsou lehčí, obratnější a doletí dále** 

DALŠÍ ČLÁNKY



**Chtěla jít bránit vlast. Nepustili ji: O slona se umíš postarat jen ty**

3. 9. 21:38 · RFE/RL

---

**NASA podruhé zrušila start rakety k měsíci. Podívejte se, jak měla letět**

3. 9. 17:54 · JAN MAREK

---



**Video zachytilo stfelce. Kdyby neselhala zbraň, je viceprezidentka po smrti**

2. 9. 11:55 · JAN MAREK




---




**Video: Takhle prý start-up oživi mamuty a tasmské tygry**

1. 9. 18:30 · JAN MAREK

- Low resource language (Czech) - not as mainstream as English
- But one can get in touch with the media company to release the data!

<b>Name</b>	dev_MLASK_visual_01-2021_06-2021.tar.gz	
<b>Size</b>	48.91 GB	
<b>Format</b>	application/x-gzip	
<b>Description</b>	dev_MLASK_visual_01-2021_06-2021	
<b>MD5</b>	a149967314ed0943287e057fefcf8008	
<a href="#">Download file</a>		
<b>Name</b>	test_MLASK_visual_07-2021_02-2022.tar.gz	
<b>Size</b>	53.22 GB	
<b>Format</b>	application/x-gzip	
<b>Description</b>	test_MLASK_visual_07-2021_02-2022	
<b>MD5</b>	70ec8735da4e9e2562cfb32a24abab76	
<a href="#">Download file</a>		
<b>Name</b>	train_MLASK_visual_02-2019_09-2019.tar.gz	
<b>Size</b>	126.79 GB	
<b>Format</b>	application/x-gzip	
<b>Description</b>	train_MLASK_visual_02-2019_09-2019	
<b>MD5</b>	63b618313b64ff0326911761732eabc2	
<a href="#">Download file</a>		



ránit vlast. Nepustili ji:  
míš postarat jen ty

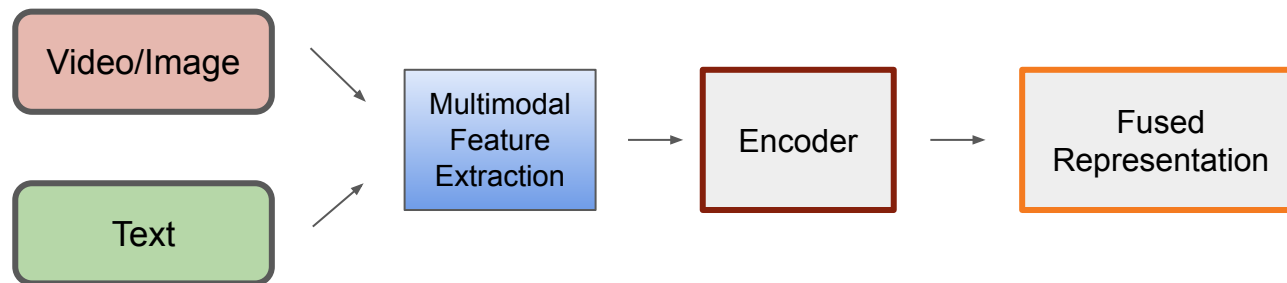
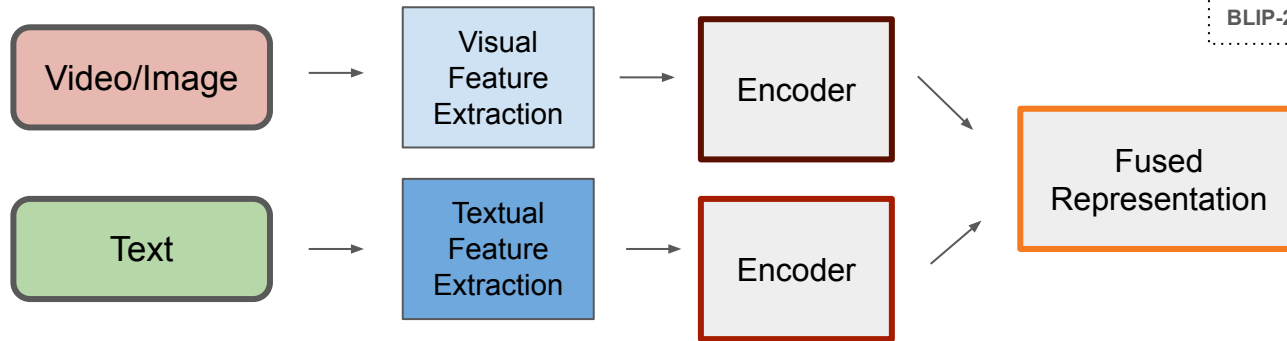
hě zrušila start rakety  
dívajte se, jak měla letět

tilo stfelce. Kdyby  
braň, je viceprezidentka po

le prý start-up oživí  
asmánské tygry

# Multimodal feature extraction and cross-modal fusion

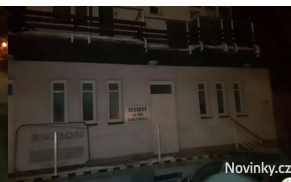
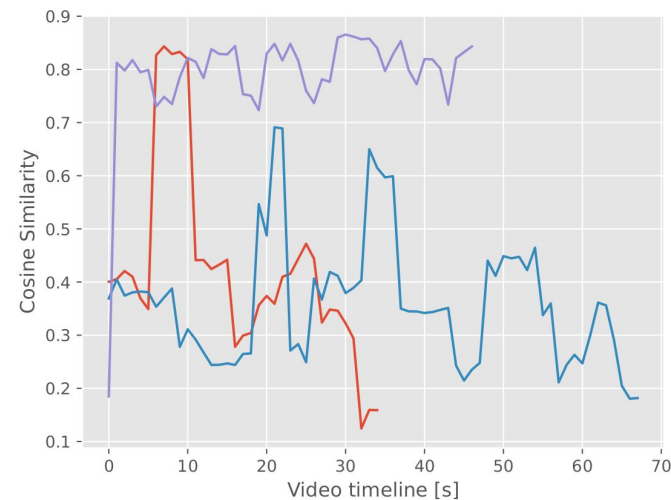
VisualBERT (Li et al., 2019),  
HERO (Li et al., 2020),  
Video-CLIP (Xu et al., 2021),  
Flamingo (Alayrac et al., 2022),  
BLIP-2 (Li et al., 2023)



# Formulation of learning signal and task-specific pre-training

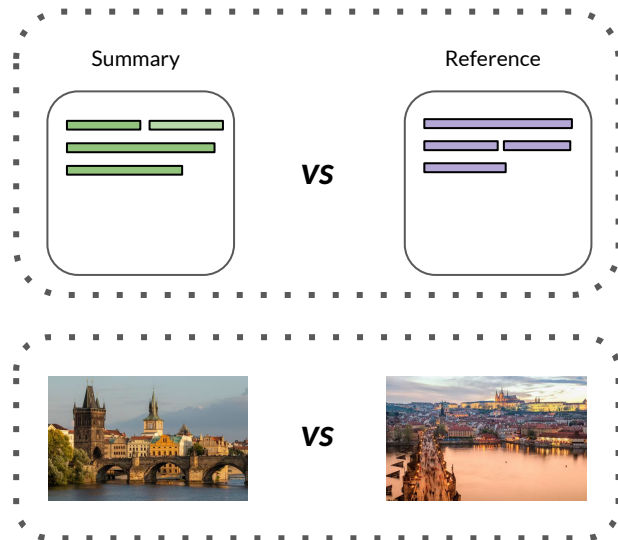
Previous works (Li et al., 2020; Fu et al., 2021) considered the most similar frame from the video as ground-truth and others as negatives.

- Single peak - ok
- Still scenes?
- Multiple peaks?

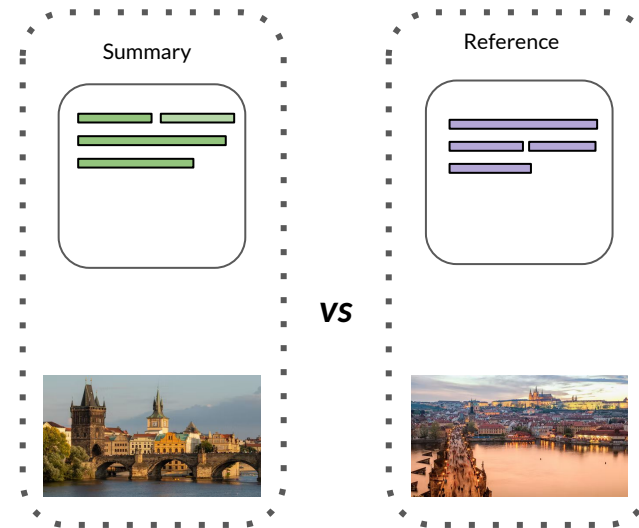


# Multimodal evaluation

Now: 🤔



Future: 😊





# Multimodal evaluation

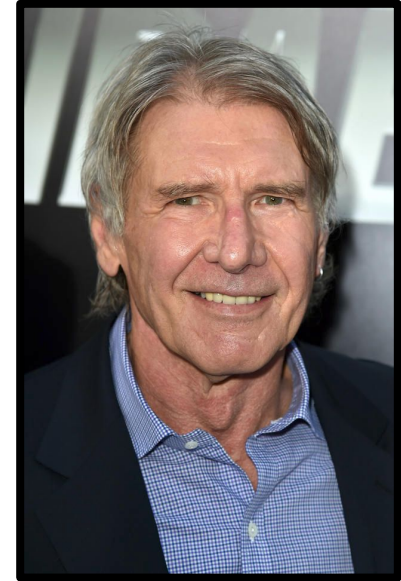
Harrison Ford **Plane Crash**  
Caused By Loose Engine Part



**Plane Crashes** - Actor  
Seriously **Injured**



Actor **Harrison Ford** **injured** in  
**crash** landing on golf course



# Multimodal evaluation

## The ship ran aground near Japan.

A cargo ship ran aground off the coast of Japan. The captain managed to free the ship, but it subsequently broke apart and began to leak fuel.



Rate 3 images on a scale of 0 to 4  
(the higher, the better).

# Multimodal evaluation

The ship ran aground near Japan.

A cargo ship ran aground off the coast of Japan. The captain managed to free the ship, but it subsequently broke apart and began to leak fuel.



	Total Score
<i>Reference</i>	$2.89 \pm 0.99$
<i>RandomV</i>	$2.39 \pm 1.15$
System A	$2.64 \pm 1.10$
System B	$2.66 \pm 1.04$

Rate 3 images on a scale of 0 to 4  
(the higher, the better).

<b># Annotators</b>	18
<b># Instances</b>	300
<b># Systems</b>	4
<b>Avg annotations per image</b>	2.54

# Institute of Formal and Applied Linguistics (ÚFAL)



- part of the Computer Science School at the Faculty of Mathematics and Physics, Charles University
- 50+ faculty members, 20+ Ph.D. students
- 150+ GPUs, 1500+ CPUs
- various resources, tools and projects
  - Universal Dependencies
  - Prague Dependency Treebank
  - UDPipe
  - LINDAT Translation
  - Charles Translator for Ukraine
  - THEaiTRE
  - ...



# References

- 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.
- 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)*. AAAI Press, 4152–4158.
- Palaskar, S., Libovický, J., Gella, S., Metze, F., 2019. **Multimodal Abstractive Summarization for How2 Videos**, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Presented at the *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, pp. 6587–6596.
- Sanabria, R., Caglayan, O., Palaskar, S., Elliott, D., Barrault, L., Specia, L., Metze, F., 2018. **How2: A Large-scale Dataset for Multimodal Language Understanding**. arXiv:1811.00347 [cs].
- Jinbae Im, Moonki Kim, Hoyeop Lee, Hyunsouk Cho, and Sehee Chung. 2021. **Self-Supervised Multimodal Opinion Summarization**. 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 388–403, Online. Association for Computational Linguistics.
- Haoran Li, Junnan Zhu, Jiajun Zhang, Xiaodong He, and Chengqing Zong. 2020. **Multimodal Sentence Summarization via Multimodal Selective Encoding**. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5655–5667, Barcelona, Spain (Online). International Committee on 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.
- Mingzhe Li, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2020. **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.
- 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.
- Tang, P., Hu, K., Zhang, L., Luo, J. and Wang, Z., 2022. **TLDW: Extreme Multimodal Summarisation of News Videos**. arXiv preprint arXiv:2210.08481.
- Mateusz Krubiński and Pavel Pecina. 2023. **MLASK: Multimodal Summarization of Video-based News Articles**. EACL 2023 Findings

# Thank you!



**krubinski@ufal.mff.cuni.cz**

**<https://github.com/ufal/MLASK>**