REFINESUMM: Self-Refining MLLM for Generating a Multimodal Summarization Dataset

Vaidehi Patil, Leonardo Ribeiro, Mengwen Liu,
Mohit Bansal, Markus Dreyer

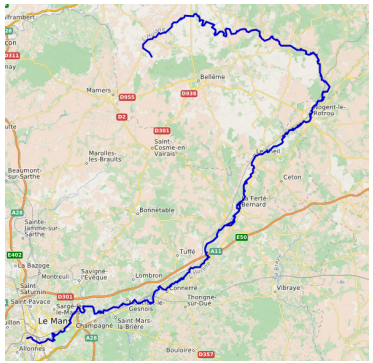
vaidehi@cs.unc.edu

Task: Multimodal Summarization

Article section

The Huisne is a 164.5 km (102.2 mi) long river in France. It is a left tributary of the river Sarthe, which it meets in Le Mans. Its source is near the town of Pervenchères, in the Orne department.
The Huisne flows through the following departments and towns:
Orne: Saint-Jouin-de-Blavou, Rémalard, Condé-sur-Huisne, Le Theil
Eure-et-Loir: Nogent-le-Rotrou
Sarthe: La Ferté-Bernard, Montfort-le-Gesnois, Le Mans.

Task: Multimodal Summarization

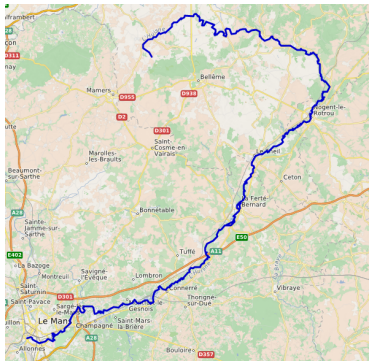
Article section	Image
The Huisne is a 164.5 km (102.2 mi) long river in France. It is a left tributary of the river Sarthe, which it meets in Le Mans. Its source is near the town of Pervenchères, in the Orne department. The Huisne flows through the following departments and towns: Orne: Saint-Jouin-de-Blavou, Rémalard, Condé-sur-Huisne, Le Theil Eure-et-Loir: Nogent-le-Rotrou Sarthe: La Ferté-Bernard, Montfort-le-Gesnois, Le Mans.	

Task: Multimodal Summarization

Article section

The Huisne is a 164.5 km (102.2 mi) long river in France. It is a left tributary of the river Sarthe, which it meets in Le Mans. Its source is near the town of Pervenchères, in the Orne department.
The Huisne flows through the following departments and towns:
Orne: Saint-Jouin-de-Blavou, Rémalard, Condé-sur-Huisne, Le Theil
Eure-et-Loir: Nogent-le-Rotrou
Sarthe: La Ferté-Bernard, Montfort-le-Gesnois, Le Mans.

Image



Self-refined summary

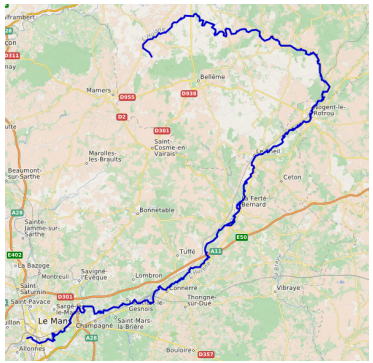
- The Huisne is a significant river in France, stretching 164.5 kilometers from its source near Pervenchères in the Orne department to its confluence with the river Sarthe in Le Mans.
- The river flows through several departments and towns, including Saint-Jouin-de-Blavou, Rémalard, Condé-sur-Huisne, and Le Theil, as well as Eure-et-Loir, Nogent-le-Rotrou, Sarthe, La Ferté-Bernhard, Montfort-le-Gesnois, and Le Mans.
- The image shows the river's path, highlighting its importance as a tributary in the region.

Task: Multimodal Summarization

Article section

The Huisne is a 164.5 km (102.2 mi) long river in France. It is a left tributary of the river Sarthe, which it meets in Le Mans. Its source is near the town of Pervenchères, in the Orne department.
The Huisne flows through the following departments and towns:
Orne: Saint-Jouin-de-Blavou, Rémalard, Condé-sur-Huisne, Le Theil
Eure-et-Loir: Nogent-le-Rotrou
Sarthe: La Ferté-Bernard, Montfort-le-Gesnois, Le Mans.

Image



Self-refined summary

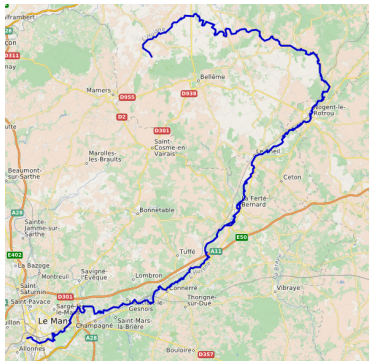
- The Huisne is a significant river in France, stretching 164.5 kilometers from its source near Pervenchères in the Orne department to its confluence with the river Sarthe in Le Mans.
- The river flows through several departments and towns, including Saint-Jouin-de-Blavou, Rémalard, Condé-sur-Huisne, and Le Theil, as well as Eure-et-Loir, Nogent-le-Rotrou, Sarthe, La Ferté-Bernhard, Montfort-le-Gesnois, and Le Mans.
- The image shows the river's path, highlighting its importance as a tributary in the region.

Task: Multimodal Summarization

Article section

The Huisne is a 164.5 km (102.2 mi) long river in France. It is a left tributary of the river Sarthe, which it meets in Le Mans. Its source is near the town of Pervenchères, in the Orne department.
The Huisne flows through the following departments and towns:
Orne: Saint-Jouin-de-Blavou, Rémalard, Condé-sur-Huisne, Le Theil
Eure-et-Loir: Nogent-le-Rotrou
Sarthe: La Ferté-Bernard, Montfort-le-Gesnois, Le Mans.

Image

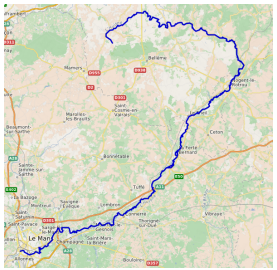


Self-refined summary

- The Huisne is a significant river in France, stretching 164.5 kilometers from its source near Pervenchères in the Orne department to its confluence with the river Sarthe in Le Mans.
- The river flows through several departments and towns, including Saint-Jouin-de-Blavou, Rémalard, Condé-sur-Huisne, and Le Theil, as well as Eure-et-Loir, Nogent-le-Rotrou, Sarthe, La Ferté-Bernhard, Montfort-le-Gesnois, and Le Mans.
- The image shows the river's path, highlighting its importance as a tributary in the region.

Challenges in Multimodal Summarization

- *Summaries generated by multimodal LLMs*
  - *lack adequate multimodal information integration*

Article section	Image	Zero-shot Summary
The Huisne is a 164.5 km (102.2 mi) long river in France. It is a left tributary of the river Sarthe, which it meets in Le Mans. Its source is near the town of Pervenchères, in the Orne department. The Huisne flows through the following departments and towns: Orne: Saint-Jouin-de-Blavou, Rémalard, Condé-sur-Huisne, Le Theil Eure-et-Loir: Nogent-le-Rotrou Sarthe: La Ferté-Bernard, Montfort-le-Gesnois, Le Mans.		• The Huisne is a 164.5-kilometer-long river in France, which is a left tributary of the river Sarthe. • It originates near Pervenchères in the Orne department and flows through Saint-Jouin-de-Blavou, Rémalard, Condé-sur-Huisne, Le Theil, Nogent-le-Rotrou, La Ferté-Bernard, Montfort-le-Gesnois, and Le Mans.

Challenges in Multimodal Summarization

- *Summaries generated by multimodal LLMs*
  - *lack adequate multimodal information integration*
  - *not faithful to the input text and image*
- *Lack of accurate and faithful datasets for the task*
  - *Getting high quality dataset is expensive (requires human annotations)*

This work

- *Self-refinement for multimodal summary generation:*
    - *Improving MLLMs using self-generated data filtered by a critic model model*

- *RefineSumm:*
    - *Dataset for multimodal summarization generated using self-refinement*

Experiment Setup

- *Dataset*
    - *Articles and Images: Obtained from randomly sampled sections and images from Wikipedia (WikiWeb2M [1])*
- *Multimodal LLM*
    - *LLaVA-v1.6-Mistral-7B [2]*

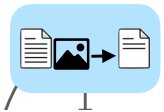
[1] Burns, Andrea, et al. "Wikiweb2m: A page-level multimodal wikipedia dataset." *arXiv preprint arXiv:2305.05432* (2023).
[2] Liu, Haotian, et al. "Llava-next: Improved reasoning, ocr, and world knowledge." (2024).

Self-refinement of MLLM for data generation

Step 1

Train a multi-dimensional critic model.

MLLM generates multi-modal summaries from article-image pairs.



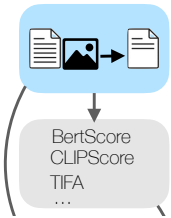
Self-refinement of MLLM for data generation

Step 1

Train a multi-dimensional critic model.

MLLM generates multi-modal summaries from article-image pairs.

Multiple automatic metrics measure the quality of each summary.



Self-refinement of MLLM for data generation

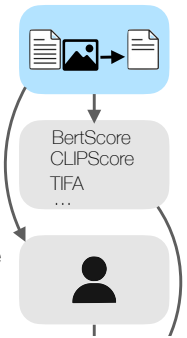
Step 1

Train a multi-dimensional critic model.

MLLM generates multi-modal summaries from article-image pairs.

Multiple automatic metrics measure the quality of each summary.

Human annotators judge multiple dimensions per summary.



Self-refinement of MLLM for data generation

Step 1

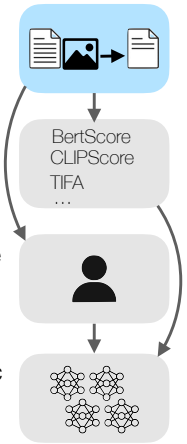
Train a multi-dimensional critic model.

MLLM generates multi-modal summaries from article-image pairs.

Multiple automatic metrics measure the quality of each summary.

Human annotators judge multiple dimensions per summary.

A multi-dimensional critic learns to map automatic to human judgements.



Self-refinement of MLLM for data generation

Step 1

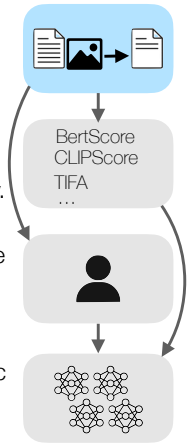
Train a multi-dimensional critic model.

MLLM generates multi-modal summaries from article-image pairs.

Multiple automatic metrics measure the quality of each summary.

Human annotators judge multiple dimensions per summary.

A multi-dimensional critic learns to map automatic to human judgements.



Step 2

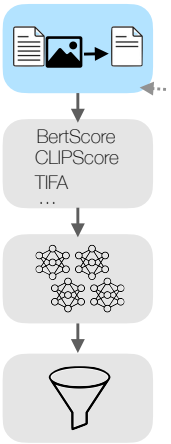
Use the critic model to filter the summaries.

MLLM generates multi-modal summaries from article-image pairs.

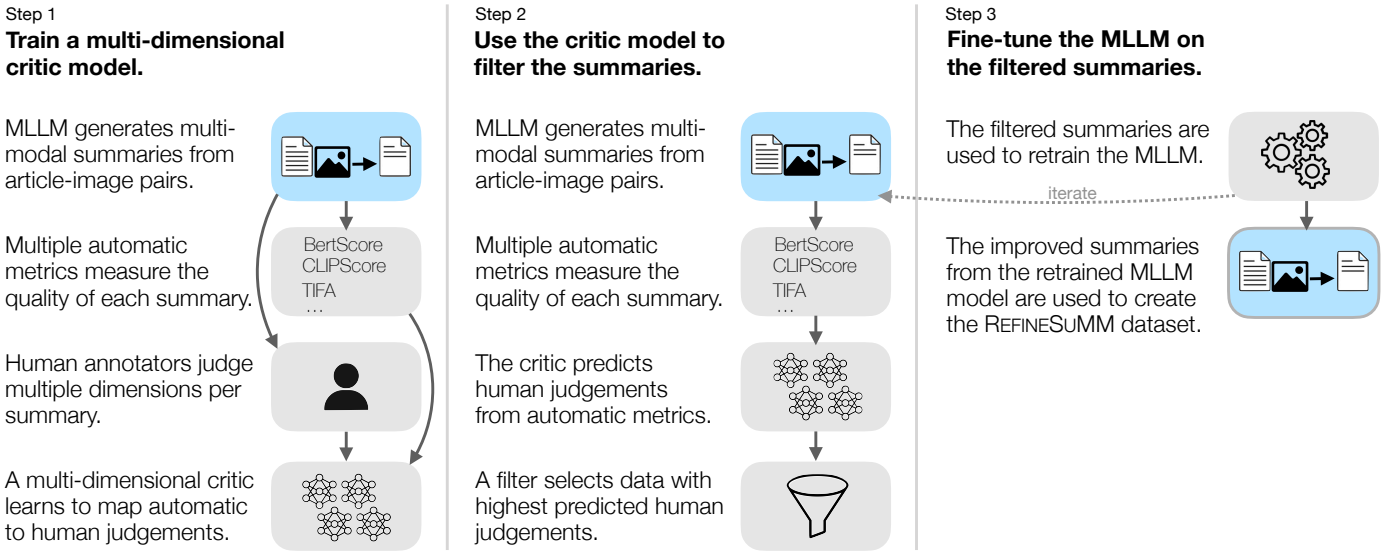
Multiple automatic metrics measure the quality of each summary.

The critic predicts human judgements from automatic metrics.

A filter selects data with highest predicted human judgements.



Self-refinement of MLLM for data generation



Experiment Setup

- *Multimodal Summary Evaluation: Automatic*
  - *BERTScore, FactCC (Faithfulness and Informativeness)*
  - *UniEval (Coherence, Consistency, Fluency, Relevance)*
  - *MINT Score (Abstractiveness)*
  - *CLIPScore (informativeness)*
  - *Inv-TIFA (faithfulness)*
  - *CLIPBERT Score (multimodal informativeness)*

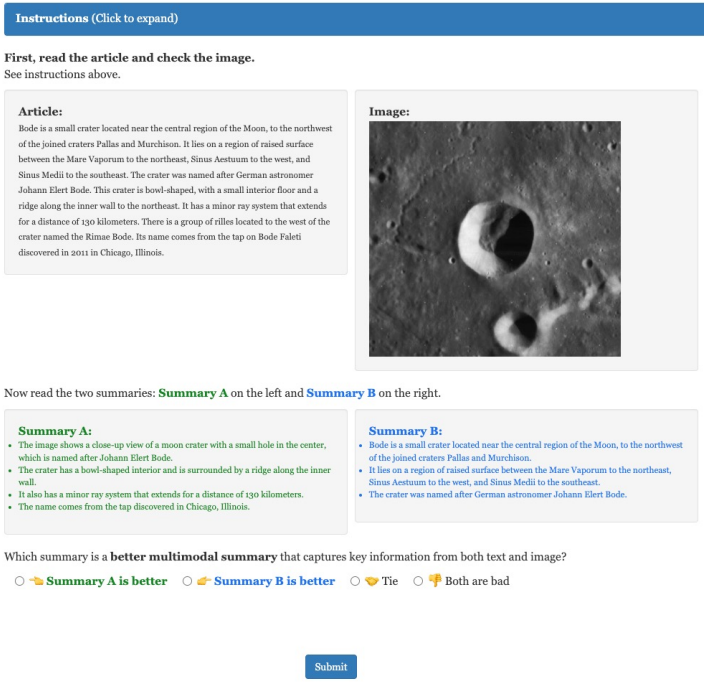
Textual metrics (w.r.t. source article)

Visual metrics (w.r.t. Image)

Multimodal
(w.r.t. both article and image)

Experiment Setup

- *Multimodal Summary Evaluation: Human*
  - *Correctness w.r.t. article*
  - *Informativeness w.r.t. article*
  - *Correctness w.r.t. image*
  - *Informativeness w.r.t. image*
  - *Multimodal summary quality (BT rating)*



Results: Automatic and Human evaluation

Modality	Metric	Zero-shot	Self-refined
Textual	BERTScore	0.894	0.879
	Overall-UniEval	0.929	0.884
	MINT	0.597	**0.630**
	FactCC	0.612	**0.636**
Visual	CLIPScore	0.284	**0.287**
	Inv-TIFA	0.610	**0.672**
Multimodal	CLIPBERTScore	0.589	0.583

Automatic Evaluation

Results: Automatic and Human evaluation

Modality	Metric	Zero-shot	Self-refined
Textual	BERTScore	0.894	0.879
	Overall-UniEval	0.929	0.884
	MINT	0.597	**0.630**
	FactCC	0.612	**0.636**
Visual	CLIPScore	0.284	**0.287**
	Inv-TIFA	0.610	**0.672**
Multimodal	CLIPBERTScore	0.589	0.583

Automatic Evaluation

Modality	Metric	Zero-shot	Self-refined
Textual	correctness	3.63	**3.67**
	informativeness	3.41	**3.43**
Visual	correctness	2.53	**3.25**
	informativeness	2.41	**3.29**
Multimodal	BT rating	1019	**1092**

Human Evaluation

Results: Automatic and Human evaluation

- More faithful summaries
- Better balance and integration of both modalities

Modality	Metric	Zero-shot	Self-refined
Textual	BERTScore	0.894	0.879
	Overall-UniEval	0.929	0.884
	MINT	0.597	**0.630**
	FactCC	0.612	**0.636**
Visual	CLIPScore	0.284	**0.287**
	Inv-TIFA	0.610	**0.672**
Multimodal	CLIPBERTScore	0.589	0.583

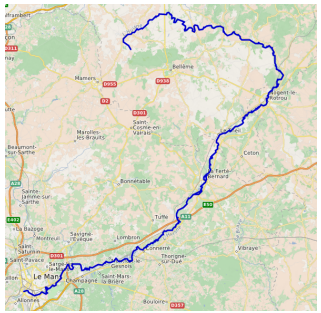
Automatic Evaluation

Modality	Metric	Zero-shot	Self-refined
Textual	correctness	3.63	**3.67**
	informativeness	3.41	**3.43**
Visual	correctness	2.53	**3.25**
	informativeness	2.41	**3.29**
Multimodal	BT rating	1019	**1092**

Human Evaluation

Results: Qualitative

➢ Better balance and integration of both modalities

Article section	Image	Zero-shot Summary	Self-refined summary
The Huisne is a 164.5 km (102.2 mi) long river in France. It is a left tributary of the river Sarthe, which it meets in Le Mans. Its source is near the town of Pervenchères, in the Orne department. The Huisne flows through the following departments and towns: Orne: Saint-Jouin-de-Blavou, Rémalard, Condé-sur-Huisne, Le Theil Eure-et-Loir: Nogent-le-Rotrou Sarthe: La Ferté-Bernard, Montfort-le-Gesnois, Le Mans.		• The Huisne is a 164.5-kilometer-long river in France, which is a left tributary of the river Sarthe. • It originates near Pervenchères in the Orne department and flows through Saint-Jouin-de-Blavou, Rémalard, Condé-sur-Huisne, Le Theil, Nogent-le-Rotrou, La Ferté-Bernard, Montfort-le-Gesnois, and Le Mans.	• The Huisne is a significant river in France, stretching 164.5 kilometers from its source near Pervenchères in the Orne department to its confluence with the river Sarthe in Le Mans. • The river flows through several departments and towns, including Saint-Jouin-de-Blavou, Rémalard, Condé-sur-Huisne, and Le Theil, as well as Eure-et-Loir, Nogent-le-Rotrou, Sarthe, La Ferté-Bernhard, Montfort-le-Gesnois, and Le Mans. • The image shows the river's path, highlighting its importance as a tributary in the region.

Results: Qualitative

- Improved faithfulness with respect to the input

❄️ The Italian wall lizard, also known as the ruin lizard, is a species of lizard native to various countries including Bosnia and Herzegovina, Croatia, France, Italy, Serbia, Montenegro, Slovenia, and Switzerland. The image shows insects.....

The Italian wall lizard or ruin lizard (Podarcis siculus, from the Greek meaning agile and feet) is a species of lizard in the family Lacertidae. P. siculus is native to Bosnia and Herzegovina, Croatia, France, Italy, Serbia, Montenegro, Slovenia, and Switzerland, but has also been introduced to Spain, ...

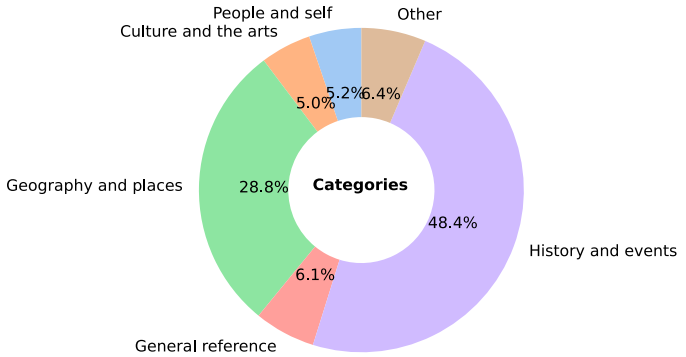
🔥 The image features a green and brown lizard with long legs, sitting on a rock. This lizard is an Italian wall lizard, which is native to various countries in Europe and has been introduced to other regions. The lizard is a habitat generalist and has a diverse diet, allowing it to thrive in different environments.

RefineSumm: Dataset statistics

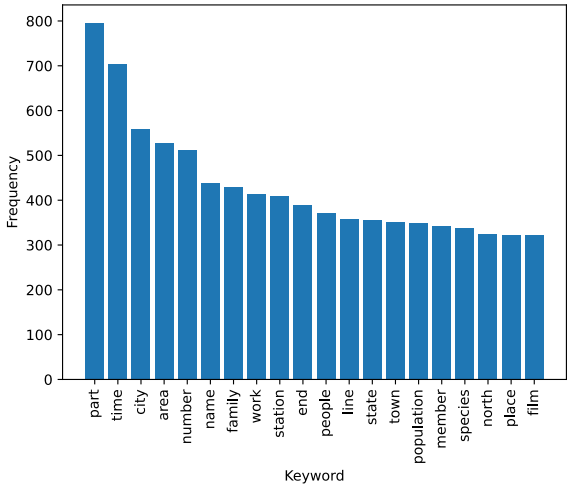
		Summary	Article
Train	Size	67024	
	Avg. Num Sent.	5.51	8.41
	Avg. Num Words	91.47	183.23
Validation	Size	4998	
	Avg. Num Sent.	5.57	8.20
	Avg. Num Words	92.99	177.50
Test	Size	4999	
	Avg. Num Sent.	5.53	8.45
	Avg. Num Words	91.84	184.57

RefineSumm: Dataset statistics

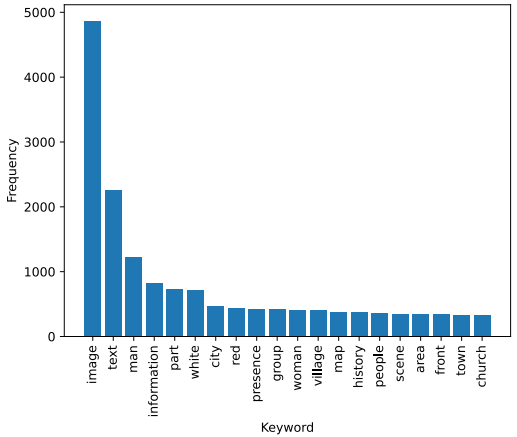
		Summary	Article
Train	Size	67024	
	Avg. Num Sent.	5.51	8.41
	Avg. Num Words	91.47	183.23
Validation	Size	4998	
	Avg. Num Sent.	5.57	8.20
	Avg. Num Words	92.99	177.50
Test	Size	4999	
	Avg. Num Sent.	5.53	8.45
	Avg. Num Words	91.84	184.57



RefineSumm: Dataset statistics



Keywords in articles



Keywords in summaries

Conclusion

- *Self-refinement pipeline for summary generation using a multimodal LLM*
- *Self-refined summaries are more faithful and involve better multimodal balance*
- *RefineSumm: Dataset for multimodal summarization using self-refinement*

Future work

- *For tasks with limited/expensive data availability:*
  - *Explore the use of self-refinement and*
  - *Critic model to learn human judgements from limited data*

Thank you!

Paper: https://www.amazon.science/publications/refinesumm-self-refining-mllm-for-generating-a-multimodal-summarization-dataset