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

Vaidehi Patil[§*] **Leonardo F. R. Ribeiro**[‡] **Mengwen Liu**[‡]
Mohit Bansal[‡,§] **Markus Dreyer**[‡]
[‡]Amazon AGI, [§]UNC Chapel Hill

Abstract

Multimodal Large Language Models (MLLMs) excel at synthesizing key information from diverse sources. However, generating accurate and faithful multimodal summaries is challenging, primarily due to the lack of appropriate multimodal datasets for fine-tuning that meaningfully integrate textual and visual modalities. To address this gap, we present a new dataset specifically designed for image-text multimodal summarization, harnessing the capabilities of state-of-the-art MLLMs. We generate summaries from Wikipedia sections and corresponding images and evaluate them across text-based, visual and multimodal dimensions, employing reference-free metrics. To refine the dataset, we: (1) filter the MLLM-generated summaries by training a critic model on human annotations and using its predictions to remove low-quality summaries; (2) fine-tune the MLLM with the filtered high-quality summaries; (3) use the fine-tuned model in turn to regenerate the summaries. This self-refinement process notably improves summary quality, as measured by human judgments and automatic multimodal metrics, resulting in a valuable dataset for multimodal summarization research.[1]

1 Introduction

In the age of information overload, efficiently extracting and summarizing key points from diverse sources is crucial. Large language models (LLMs) have become powerful tools, generating human-like text with impressive fluency. However, their limitations in faithfully capturing information become evident when faced with diverse, heterogeneous multimodal inputs like text and images. This poses a significant challenge for the task of multimodal summarization, which requires condensing rich information from multiple modalities into concise, coherent summaries capturing the essence of the original content. The task has far-reaching applications, from enhancing content browsing, information access and retrieval, to promoting accessibility for diverse needs.

Figure 1: The summary generated from zero-shot manner (snowflake icon) has incorrect information (in red) like insects and is not able to ground that information with the text which talks about lizard whereas the summary generated by our self-refined model (fire icon) has image information (in green) that has been coherently combined with the information from the input article.

Multimodal Large Language Models (MLLMs) utilize the strengths of LLMs to excel in handling multimodal data and offer a new research paradigm that can potentially address these challenges. Jangra et al. (2023) highlight the challenges faced by multimodal summarization such as identifying key information from both modalities and using it to generate coherent and faithful summaries. Generating content that faithfully reflects the original information is especially challenging when dealing with complex, heterogeneous multimodal inputs (Jing et al., 2023; Wan and Bansal, 2022). To illustrate this point, Figure 1 shows a summary generated by a multimodal LLM that is not fully faithful to the input.

High-quality datasets for multimodal summariza-

*Work done as an intern at Amazon AGI.

[1]The dataset is publicly available at https://github.com/amazon-science/refinesumm.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13773–13786
August 11-16, 2024 ©2024 Association for Computational Linguistics

tion that demonstrate faithful and informative summaries could guide the models towards improving fidelity of their generations. Such datasets should comprehensively assess cross-modal capabilities, ensuring the coherent and faithful integration of information from both textual and visual sources in the summaries. However, such datasets are not available today; existing multimodal datasets (Wan and Bansal, 2022; Burns et al., 2023) are often constructed by taking the first sentence as the summary, leading to short summaries where the image plays a limited role. This underscores the need for improved datasets that truly challenge the cross-modal capabilities of such models while ensuring the meaningful integration of both modalities in the summarization process.

To address this need, we present a new benchmark dataset for image-text multimodal summarization, featuring a key innovation: a *self-refinement* process for creating high-quality summaries. In this process, we first evaluate the summaries generated by an MLLM along textual, visual, and multimodal dimensions and then filter them systematically with the help of a set of classifiers trained on data obtained from human annotations. These classifiers serve as a critic that predicts human judgments from automatic metrics. The high-quality summaries that pass the critic-based filter are then used to fine-tune the MLLM. This fine-tuned, self-refined MLLM then generates the summaries that constitute our dataset.

Textual and visual faithfulness of the generated summaries, as measured by automatic evaluation metrics improves after refining the MLLM generator. Human evaluation results show that the quality of summaries generated by our approach along dimensions such as informativeness and correctness with respect to both text and image improves after refinement. Our dataset creation method is cost-effective, using minimal human annotation for training the critic model to screen the summaries.

In essence, we present a cost-effective approach for constructing a high-quality multimodal summarization dataset to benefit further research and development in this field. Overall, our main contributions are as follows:

- **Self-refinement of MLLM**: We develop a self-refinement process that generates, evaluates, and filters summaries generated by a multimodal LLM, and employs the filtered summaries to fine-tune the MLLM, leading to a high-quality dataset while minimizing human annotation costs. We show that the proposed refinement process improves the quality of the generated summaries along textual and visual dimensions.

- **Multimodal Summarization Dataset**: We release the resulting summaries as a dataset for multimodal summarization, called REFINESUMM, containing 77k article-image input pairs and corresponding summaries.

2 Related Work

2.1 Multimodal Summarization Datasets

Multimodal data integration for text summarization has shown promising results in enhancing summary quality. Liu et al. (2023a) introduce a visual instruction tuning dataset consisting of multiple tasks for mitigating in hallucinations MLLM output. However, their dataset does not include the task of multimodal summarization.

Li et al. (2017, 2018); Palaskar et al. (2019) have delved into the integration of multimodal data, including video and audio transcripts, to augment textual documents with the goal of enhancing the quality of textual summaries. Zhu et al. (2018), who introduced the Multimedia Summarization of Media Objects (MSMO) task, developed a model that jointly generates text and selects the most pertinent image from a predetermined set of images. Li et al. (2020) and Fu et al. (2021) were the first to address the Video-based Multimedia Summarization of Media Objects (VMSMO) problem.

Subsequently, in the follow-up work by Tang et al. (2023), video-article pairs were summarized by condensing each pair into a single frame and a one-sentence summary, achieved through an optimal transport-based unsupervised training strategy. Krubiński and Pecina (2023) recently introduced a Czech language dataset comprising video-based news articles, each accompanied by a textual summary and a cover picture. Li et al. (2018) construct a multimodal image-text summarization dataset based on an annotated corpus. However, their input text consists of just one sentence and the summary is even shorter. Moreover, image information for constructing summaries remains largely unexplored in these works.

Burns et al. (2023) introduce the task of section summarization for the WikiWeb2M dataset, employing the first sentence of the section as the ground truth summary. However, such sentences often do not comprehensively capture salient information from image and text modalities. To the best of our knowledge, this paper presents the first work that uses MLLMs for curating an image-text multimodal summarization dataset.

Models for Multimodal Summarization. Multimodal fusion for summarization has seen growing interest, aiming to achieve more comprehensive and informative summaries by incorporating diverse data types. Khullar and Arora (2020) introduce a summarization model which can combine three modalities: text, audio and video. He et al. (2023) introduce a unified video-text summarization framework that attends to and aligns the two modalities while leveraging their time correspondence and returns the important frames and sentences as the summary. Yu et al. (2021) develop a method to inject vision information in text-only generative pre-trained LMs for the task of multimodal abstractive summarization. However, they focus on combining diverse modalities like video and audio, or rely on complex architectures. Ghosh et al. (2023) employ the combination of CLIP and a general purpose LLM for the task of multimodal question summarization. However, the application of LLMs for image-text summarization remains less explored. Our work addresses this gap by employing state-of-the-art pre-trained multimodal LLMs specifically for image-text summarization. This approach allows us to directly leverage the strengths of MLLMs in capturing and reasoning about both visual and textual information, reducing the need for elaborate fusion mechanisms and leading to the creation of a high-quality image-text summarization dataset.

2.2 LLMs for Dataset Creation

The application of LLMs for data synthesis has gained significant traction recently, offering a promising avenue to address data scarcity challenges in various tasks. Several studies demonstrate the effectiveness of LLMs in generating synthetic data for specific domains.

Su et al. (2023) leverage LLMs to generate target-domain text corpora, enhancing the performance of Automatic Speech Recognition systems in specific domains. Rosenbaum et al. (2022b) introduce a technique for creating annotated data for intent classification and slot tagging tasks by fine-tuning the AlexaTM 5B model. Rosenbaum et al. (2022a) utilize AlexaTM 20B model to generate synthetic data that supplements the training set for a smaller model, achieving a 40x reduction in model size.

Our work builds upon these advancements by exploring the use of an MLLM for data synthesis in the context of multimodal summarization. This goes beyond existing approaches by: (1) Focusing on multimodal data: We address the specific challenge of generating summaries incorporating both text and image information, (2) Self-refinement: We propose a refinement process involving fine-tuning the LLM with high-quality summaries, leading to an improvement in data quality.

2.3 Self-Refinement

Madaan et al. (2023) introduce self-refinement that uses feedback from an initial output to refine itself and generate its subsequent self-refined output. Chen et al. (2023b) utilize the self-refinement framework to self-debug for the task of code generation. Gulcehre et al. (2023) show that Reinforced Self-Training, inspired by growing batch reinforcement learning, improves the quality of LLMs by generating offline training data aligned with human preferences, ultimately enhancing performance in tasks like machine translation with efficiency.

Chen et al. (2024) propose Self-Play Fine-Tuning (SPIN), a method using a self-play mechanism to train and improve LLMs without additional human annotations. The self-play mechanism involves an LLM refining its capability by playing against instances of itself.

In this work, we create a dataset by self-improving summaries generated using the multimodal LLM LLaVA-v1.6-Mistral-7B (Liu et al., 2024; Jiang et al., 2023), with careful feedback generated from a critic model trained on data obtained from human annotations.

3 REFINESUMM Dataset

3.1 Problem Formulation

In multimodal summarization, we are given an article $\mathcal{L}$ containing textual content and an accompanying image $\mathcal{V}$ relevant to the text in $\mathcal{L}$; we aim to generate a concise summary $\mathcal{S}$ that coherently integrates and faithfully reflects the key information from both $\mathcal{L}$ and $\mathcal{V}$.

3.2 Data Collection

To construct a dataset suitable for image-text summarization, we leverage WikiWeb2M (Burns

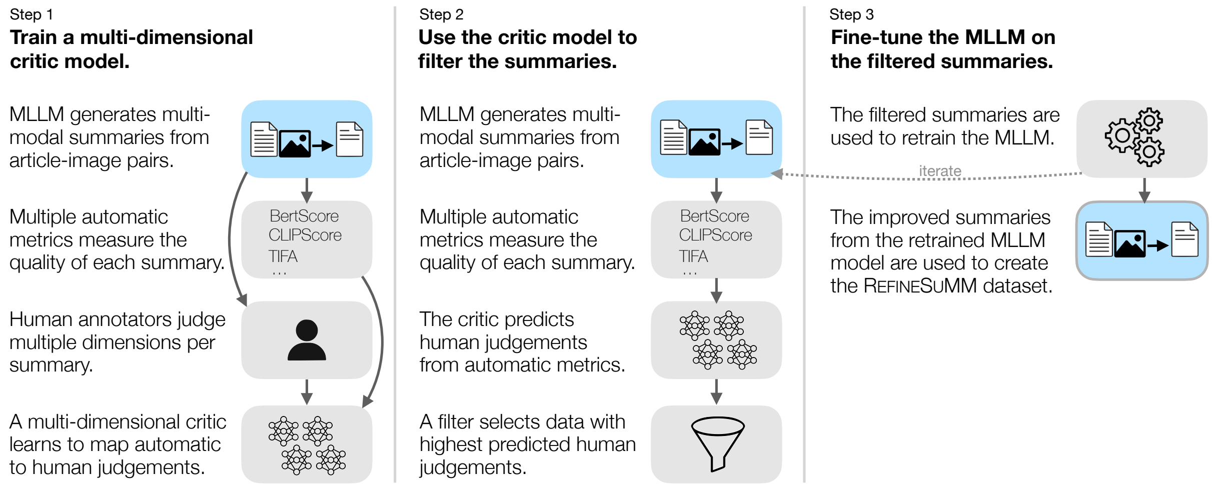

Figure 2: This figure describes the three high-level steps required to create our REFINESUMM dataset. The three steps are described in Sections 3.3.1 through 3.3.3.

et al., 2023), a large-scale dataset scraped from Wikipedia. We randomly sample a representative set of 80k samples for our task. Since we focus on the task of image-text summarization, each sample consists of a pair ($\mathcal{L}_i$, $\mathcal{V}_i$), where $\mathcal{L}_i$ represents the text extracted from a specific section within a Wikipedia article, and $\mathcal{V}_i$ denotes the corresponding image in the section that is visually associated with the section's content.

3.3 Dataset Generation

Figure 2 presents an an overview of our proposed method for dataset generation. It shows three steps, which we describe below.

3.3.1 Step 1: Train a multi-dimensional critic model

The first step is to train a multi-dimensional critic model. This requires generating summaries, scoring them using automatic metrics, collecting human judgments and training the critic. We now describe these tasks in detail, aligned with Figure 2.

MLLM generates multimodal summaries from article-image pairs. We employ a state-of-the-art MLLM, LLaVA-v1.6-Mistral-7B, to generate the summaries for the collected samples. The model receives both article $\mathcal{L}_i$ and the associated image $\mathcal{V}_i$ as input and is prompted to generate a concise and coherent summary $\mathcal{S}_i$. To ensure the brevity of summaries compared to the input text, we filter out summaries $\mathcal{S}_i$ exceeding the length of the original text $\mathcal{L}_i$.

Multiple automatic metrics measure the quality of each summary. The generated summaries are then evaluated by multiple automatic metrics across textual, visual and multimodal dimensions of summary quality. We describe the metrics used in Section 4.

Human annotators judge multiple dimensions per summary. In addition to the automatic metrics, we collect human judgments of summary quality. Using Amazon Mechanical Turk, we recruited a diverse pool of annotators to evaluate a representative subset of 3,000 validation set summaries. Details of this setup are described in Section 4.3. For each sample ($\mathcal{L}_i$, $\mathcal{V}_i$, $\mathcal{S}_i$), annotators provided ratings on four key dimensions, each on a scale of 1 to 4: (1) Correctness of $\mathcal{S}_i$ with respect to text $\mathcal{L}_i$, (2) Informativeness of $\mathcal{S}_i$ with respect to text $\mathcal{L}_i$, (3) Correctness of $\mathcal{S}_i$ with respect to image $\mathcal{V}_i$, (4) Informativeness of $\mathcal{S}_i$ with respect to image $\mathcal{V}_i$.

Using informativeness, we quantify whether the summary contains enough relevant information from the respective modality. Using correctness, we quantify the faithfulness of the generated summary with respect to both the article and the image.

A multi-dimensional critic learns to map automatic to human judgments. Finally, we train a multi-dimensional critic model, consisting of four classifiers, to predict human judgments in a scalable manner. In particular, each of the classifiers separately learns to map all the automatic metrics to one of the four dimensions of human judgments described above. Human ratings on a 1-4 scale are transformed into binary labels for each annotation dimension (correctness and informativeness with respect to text/image). Ratings 1 and 2 are mapped to class 0 (Low Quality), while ratings 3 and 4 become class 1 (High Quality). This process simplifies the task for the critic model while retaining key quality indicators. To ensure robust model training and evaluation, the annotated subset is further divided into training (2400 samples) and validation sets (600 samples). This split enables assessment of the critic model's generalization ability. We employ four independent Multi-Layer Perceptrons (MLPs), each dedicated to predicting the binary class label (0 or 1) for one annotation dimension. The automatic evaluation metrics are used as input features for the MLPs.

3.3.2 Step 2: Use the Critic Model to Filter the Summaries

After the multi-dimensional critic model is trained, we use it to predict human judgments along the four dimensions of correctness and informativeness with respect to both text and image. For each sample and each annotation dimension, the trained critic models predict the probability of belonging to class 1 (High Quality). These probabilities offer a continuous measure of confidence in the summary's quality for each dimension. We employ individual thresholds for each dimension to selectively retain high-quality summaries. These thresholds are determined through systematic hyperparameter tuning over a range of values between 0.1 and 0.9 with an interval of 0.1 (see Table 8). This grid search ensures we find the optimal thresholds that distinguish high-quality summaries while making sure that a fixed number of summaries pass the thresholds. Table 6 presents the precision of the critic model at the chosen threshold values on a held-out validation set. To ensure comprehensive quality assessment, we apply all four dimension-specific filters. Only summaries that pass all criteria are retained. This multimodal filtering ensures that the final set of summaries exhibits high quality across various aspects, encompassing both textual and visual faithfulness and correctness.

3.3.3 Step 3: Fine-Tune the MLLM on the Filtered Summaries

Following the identification of high-quality summaries in Step 2, we use these examples to enhance the capabilities of the LLaVA-v1.6-Mistral-7B model through fine-tuning. We fine-tune the multimodal projector module of the LLaVA model, using the filtered high-quality summaries from the training set. This fine-tuning process allows the model to learn from these exemplars, improving its ability to generate summaries that align with human preferences and quality standards.

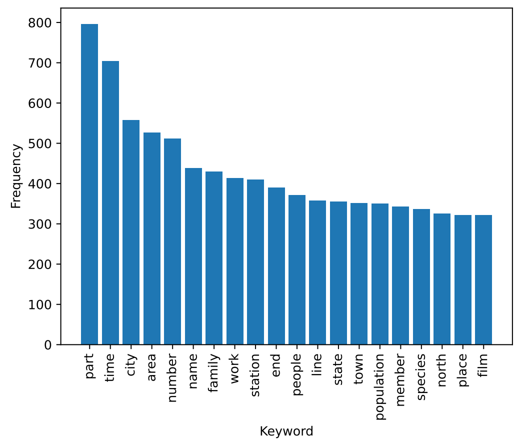

Figure 3: Most commonly occurring keywords in the articles in our REFINESUMM dataset. Many of them are related to time and location.

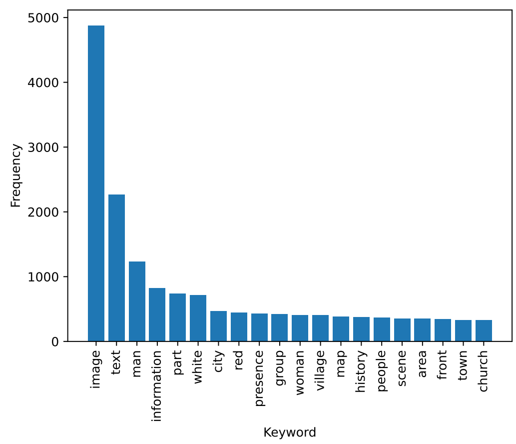

Figure 4: Most commonly occurring keywords in the summaries in REFINESUMM. Many of them are related to commonly occurring objects in images and locations.

Self-Refinement. Following the fine-tuning step, we use the newly enhanced LLaVA-v1.6-Mistral-7B model to generate new summaries that constitute our REFINESUMM dataset. Using stratified random splitting, we split the dataset into train, validation, and test sets of 67k, 5k, and 5k samples, see Table 1.

3.4 Dataset Analysis

Table 1 shows the dataset sizes and statistics across the three splits in the data.

We observe that the generated multimodal summaries are considerably shorter than the original text, with an average length of approximately one-half the article length (see Table 1). Despite their conciseness, these summaries incorporate information from both textual and visual modalities, demonstrating the model's ability to extract key content and present it in a succinct manner.

Figure 3 shows the most commonly occurring words in the articles, many of which are related to factual content such as time and location. Figure 4 shows the most commonly occurring words in the summaries and indicates that the generated summaries predominantly feature words associated with frequently occurring objects within the images and locations reflecting an emphasis on visually salient content.

Figure 5 shows the distribution of categories of articles in REFINESUMM. Many of the articles fall in the categories of history and events, and geography and places, which also explains the high frequencies of terms such as "time" and "years", "area" and "city" in Figure 3.

4 Experimental Setup

We now describe the evaluation metrics that we use for automatic reference-free evaluation of the generated summaries in Step 2. The automatic assessment scores are used as input features of the critic model trained in Step 2 of the proposed self-refinement pipeline.

4.1 Textual Metrics

We assess the internal quality of the summaries using coherence, consistency, fluency and relevance from the UniEval framework (Zhong et al., 2022). While coherence, consistency and fluency are reference-free metrics, when computing relevance score, we use the input text which serves as a proxy for reference summaries. Following previous work (Ribeiro et al., 2023), we use FactCC (Kryscinski et al., 2020) and BERTScore (Zhang et al., 2020) for evaluating the summary faithfulness to the input text. While BERTScore measures similarity, research shows that it is correlated with

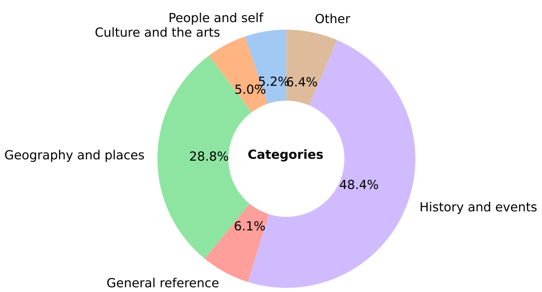

Figure 5: Categories of the documents in REFINESUMM.

		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

Table 1: Number of ($\mathcal{L}$, $\mathcal{V}$, $\mathcal{S}$) samples in each split of the dataset. Average number of words and sentences in both articles and summaries in each split. Summary length is approximately one-half of the article length.

human judgments of faithfulness (Fischer et al., 2022). We also evaluate the MINT score introduced by Dreyer et al. (2023) that quantifies the degree of abstractiveness of summaries based on textual overlap between input articles and summaries.

4.2 Visual Metrics

We employ CLIPScore (Hessel et al., 2021) for assessing the text-image compatibility, indicating how well the summary captures information from the image. We compute CLIPScore sentence-wise and average the scores for a final measure of compatibility score between the summary and the image.

Inspired by the Text-to-Image Faithfulness Evaluation (TIFA) metric (Hu et al., 2023), we propose its extension, Inverse TIFA (Inv-TIFA), for evaluating fine-grained faithfulness of a summary with respect to the input image. For each sentence in the summary, we generate questions along with their answers based on the summary sentence. We then feed the questions along with the image to a VQA model and assess the accuracy of the answers predicted from the image as compared to the answers predicted from the summary sentence (see Appendix A for more details about Inv-TIFA computation). We average this accuracy score across all sentences to get the faithfulness score of the summary with respect to the image. Finally, we employ CLIPBERTScore (Wan and Bansal, 2022) to assess multimodal faithfulness.

4.3 Human Evaluation

We conduct human evaluations using Amazon Mechanical Turk to determine how informative and faithful to the source text and image (correctness) the generated summaries are. We also conduct pairwise preference annotation of multimodal summaries.

In order to ensure high-quality human judgment, we use several mitigation strategies such as simplified task setups, clear annotation guidelines, task-specific qualification tests, and time checks to exclude potential spammers. We give annotators fair compensation. Figure 6 shows the template for evaluating the accuracy and inconsistency of the summary with respect to the input document and image, while Figure 7 presents the template for the annotation of pairwise preference of multimodal summaries. We also use a bonus incentive structure. Every worker who passes the automatic quality checks receives a bonus at the end. In addition, we only consider workers from a country whose main language is English, who has completed 100 or more HITs so far with an acceptance rate of 95% or higher.

We adopt the Bradley–Terry (BT) model[2] for deriving scores from the preference annotations, in light of recent findings (Boubdir et al., 2023) indicating that the choice of hyperparameters and comparisons makes a significant difference in ELO scores. Using a maximum likelihood estimator (MLE), BT estimates the underlying ELO model with a fixed but unknown pairwise win rate.

We sample 500 summaries for each configuration evaluated. For each summary (or pair of summaries), we collect scores from 3 annotators and compute the final score using majority voting.

4.4 Self-Refinement

We perform all experiments on a subset of the WikiWeb2M dataset, which we process as described in Step 1 of Section 3.3. In the proposed self-refinement approach, we fine-tune on a filtered subset of the train split. We report the results of the evaluation on the test split of REFINESUMM in Table 2.

5 Results

Through our experiments, we aim to answer the following questions:

5.1 Does the self-refinement of MLLMs result in better multimodal summaries?

We evaluate the summaries generated after self-refining the model along the text-based, visual and multimodal dimensions discussed in Section 4 and list the results in Table 2.

[2]https://lmsys.org/blog/2023-12-07-leaderboard/

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

Table 2: Automatic evaluation of summaries generated zero-shot manner and self-refinement.

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**

Table 3: Human evaluation of summaries generated by LLaVA-v1.6-Mistral-7B before (Zero-shot) and after self-refinement (Self-refined). The summaries produced by the self-refined model exhibit a better balance between the two modalities, with a stronger focus on the visual modality, as evidenced by the increase in informativeness and accuracy with respect to the image (Section 5.1).

We observe that fine-tuning using the filtered summaries results in a substantial improvement of 2.4% in FactCC, which evaluates faithfulness of the summaries with respect to the input text. We also observe a substantial improvement of 6.2% in the Inv-TIFA score, which indicates that the self-refinement approach helps improve visual faithfulness to the generated text. We also observe a net increase of 0.3% in CLIPScore, which indicates that the approach encourages summaries that effectively fuse information from both modalities without ignoring the visual modality. Self-refinement makes the summaries more abstractive as indicated by 3.3% increase in the MINT score.

We also perform human evaluations to evaluate whether the multimodal summary quality improves after refinement. We observe that both correctness and informativeness with respect to the images improve notably by 0.72 and 0.88 points out of 4, while informativeness and correctness with respect to the articles slightly improve as well. Thus, the resulting summaries have a better balance of information from both modalities.

5.2 Can the self-refinement process be used to iteratively enhance the summaries?

We investigate whether performing another iteration of the self-refinement procedure by using the fine-tuned model as the base model in the approach leads to any further improvements in the generated summaries. We observe that after another iteration, the number of summaries that make their way through the screening filter rises, indicating improvement in the quality of generated summaries.

We observe that the nature of summaries becomes more inclined towards text and less inclined towards image after the second round of finetuning as evident from the drop in visual scores Table 4. However, the automatic evaluations in Table 7 do indicate text-specific improvements in the summaries in the second iteration compared to the first iteration. While this second round of fine-tuning helps achieve summaries that are more informative of the textual modality, they are not as balanced as the first iteration of self-refined summaries.

5.3 Case Study

To supplement our quantitative results, we present qualitative examples of summaries before and after self-refinement in Table 5. We show that while zero-shot summaries contain negligible image information, the self-refined summary effectively incorporates relevant image content, resulting in a more coherent and comprehensive representation of the source material. This enhancement is generally evident in the improved alignment of textual descriptions with visual elements, demonstrating the capability of self-refinement to bridge the gap between textual and visual data. Furthermore, refined summaries exhibit increased fluency and contextual relevance, highlighting the potential of self-refinement techniques to improve the overall quality of generated summaries.

More specifically, in the first example, the fine-tuned summary provides a more accurate and complete representation of the information given in the article section. It mentions the significance of the river, its confluence with the river Sarthe, and adds contextual information on the significance of the river as a tributary in the region. The fine-tuned summary better describes how the image relates to the text by explicitly mentioning that the image shows the path of the river. This direct connection between the image and the text improves the reader's understanding and visualization of the information provided. In contrast, the zero-shot summary lacks completeness and context and does not highlight the importance of the river or relate the image to the text effectively.

Similarly, in the second example, the summary generated by the proposed approach includes more relevant details from the article. It better describes the key factors of the input article while also effectively representing the image. It mentions the large, old stone church with a prominent tower and clock, directly tying the visual aspects of the image to the textual description. In contrast, the zero-shot summary omits specific architectural details (tower) and provides a more generic overview that does not as effectively capture the essence of both the image and the article.

Modality	Metric	Iter 1	Iter 2
Textual	correctness	3.67	3.45
	informativeness	3.43	3.30
Visual	correctness	3.25	2.58
	informativeness	3.29	2.52
Multimodal	BT rating	1092	889

Table 4: Human evaluation-based comparison between the first iteration of self-refinement and subsequent modality-specific second iterations. The subsequent iterations helps achieve summaries that are more informative of the respective modalities less balanced than first iteration self-refined summaries (Section 5.2).

6 Conclusion

We have introduced a self-refinement pipeline utilizing a multimodal LLM, which we apply to create a new dataset for multimodal summarization of image-text pairs. Within this pipeline, we train a model that serves as a critic that predicts human judgments from automated metrics. We use this multi-dimensional critic to identify high-quality summaries. Our analysis shows that the proposed self-refinement approach yields improvements in the quality of generated summaries. This improvement is evident in improved faithfulness to both individual modalities (text and image) and better fusion of multimodal information, as demonstrated by human evaluation. Furthermore, we show a case study based on examples of summaries before and after self-refinement which suggests that self-refinement can lead to a better balance of modalities in the generated summaries. Thus, this simple and cost-effective dataset generation procedure offers considerable potential for tasks where human

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.
Both village and church are said to have been founded by the Archbishop of Canterbury, St Dunstan, in 960 CE, and there are legends surrounding his connection with the village. Dunstan is supposed to have become an ironworker and run a small forge next to the church. ... The Anglican church is dedicated to Dunstan. In 1389 much of the village and most of the church were destroyed by fire; the latter was struck by lightning in the 17th century. The church was subsequently rebuilt in the fifteenth century. The church is in a mostly Perpendicular style and has a s̈quat, shingled broach spire.̈ Inside the church there are a number of graves made of iron for the families of Mayfield's ironmasters. There are three other religious buildings in Mayfield: the Roman Catholic church of St Thomas of Canterbury (in Station Road), Coggins Mill Church (a Free Church, also in Station Road), and Mayfield Baptist Chapel (on South Street).		• The Anglican church in Mayfield is dedicated to St Dunstan, who is said to have founded the village and church in 960 CE. • The church, which is in a Perpendicular style, has a s̈quat, shingled broach spireänd features a number of graves made of iron for the families of Mayfield's ironmasters. • The village and church were destroyed by fire in 1389, and the church was rebuilt in the fifteenth century. • There are also three other religious buildings in Mayfield: the Roman Catholic church of St Thomas of Canterbury, Coggins Mill Church, and Mayfield Baptist Chapel.	• The image shows a large, old stone church with a prominent tower and a clock. • The church is dedicated to St Dunstan, an ironworker who is said to have founded the village and church in 960 CE. • The church has a rich history, with legends surrounding its connection with the village. • The church is in a mostly Perpendicular style and has a "squat, shingled broach spire." • Inside the church, there are a number of graves made of iron for the families of Mayfield's ironmasters. • The church is situated in a village, with other religious buildings in the area, including a Roman Catholic church of St Thomas of Canterbury, a Free Church, and a Baptist Chapel.

Table 5: Example of summaries generated in a zero-shot fashion and using the refined model from an article section and an image. Article sentences that are not mentioned in summaries are replaced with ellipsis. The zero-shot summaries fail in representing the image, whereas the summaries of the self-refined model better describe the key factors of the input article while also describing the image.

annotation is a significant bottleneck.

7 Limitations

Language models have the potential to magnify existing biases present in the data and may generate text that contains factual inaccuracies. The multimodal LLMs used in this paper are not exempt from such challenges. Addressing these shortcomings is a subject of ongoing research. We see our refinement technique, which uses a multimodal faithfulness and informativeness critic model based on human judgments, as a step toward more accurate and inclusive models. Additionally, licensing restrictions imposed another form of limitation on this work, as results with MiniGPT-v2 (Chen et al., 2023a) and other models based on Llama 2 were excluded due to these constraints.

The procedures of model fine-tuning and automatic scoring, especially Inverse TIFA are computationally expensive. To manage these demands efficiently, we decided to sample 80k examples from the full WikiWeb2M dataset, which contains 2M examples, to construct our dataset.

Finally, we recognize the importance of ethical considerations in the development and deployment of language models. A detailed discussion of these ethical aspects is provided in Appendix A.1.

References

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. 2023. Elo uncovered: Robustness and best practices in language model evaluation.

Andrea Burns, Krishna Srinivasan, Joshua Ainslie, Geoff Brown, Bryan A Plummer, Kate Saenko, Jianmo Ni, and Mandy Guo. 2023. A suite of generative tasks for multi-level multimodal webpage understanding. *arXiv preprint arXiv:2305.03668*.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023a. MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2023b. Teaching large language models to self-debug. *arXiv preprint arXiv:2304.05128*.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*.

Markus Dreyer, Mengwen Liu, Feng Nan, Sandeep Atluri, and Sujith Ravi. 2023. Evaluating the tradeoff between abstractiveness and factuality in abstractive summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2044–2060.

Tim Fischer, Steffen Remus, and Chris Biemann. 2022. Measuring faithfulness of abstractive summaries. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 63–73.

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.

Akash Ghosh, Arkadeep Acharya, Raghav Jain, Sriparna Saha, Aman Chadha, and Setu Sinha. 2023. Clipsyntel: Clip and llm synergy for multimodal question summarization in healthcare. *arXiv preprint arXiv:2312.11541*.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. 2023. Reinforced self-training (ReST) for language modeling. *arXiv preprint arXiv:2308.08998*.

Bo He, Jun Wang, Jielin Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. 2023. Align and attend: Multimodal summarization with dual contrastive losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14867–14878.

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

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *arXiv preprint arXiv:2303.11897*.

Anubhav Jangra, Sourajit Mukherjee, Adam Jatowt, Sriparna Saha, and Mohammad Hasanuzzaman. 2023. A survey on multi-modal summarization. *ACM Computing Surveys*, 55(13s):1–36.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, and Xinya Du. 2023. FAITHSCORE: Evaluating hallucinations in large vision-language models. *arXiv preprint arXiv:2311.01477*.

Aman Khullar and Udit Arora. 2020. MAST: Multimodal abstractive summarization with trimodal hierarchical attention. In *Proceedings of the First International Workshop on Natural Language Processing Beyond Text*, pages 60–69.

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

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. 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 Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4152–4158. International Joint Conferences on Artificial Intelligence Organization.

Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2017. Multi-modal summarization for asynchronous collection of text, image, audio and video. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1092–1102, Copenhagen, Denmark. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

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.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 1(2):9.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

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.

Leonardo F. R. Ribeiro, Mohit Bansal, and Markus Dreyer. 2023. Generating summaries with controllable readability levels. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11669–11687, Singapore. Association for Computational Linguistics.

Andy Rosenbaum, Saleh Soltan, Wael Hamza, Amir Saffari, Marco Damonte, and Isabel Groves. 2022a. CLASP: Few-shot cross-lingual data augmentation for semantic parsing. In *AACL-IJCNLP 2022*.

Andy Rosenbaum, Saleh Soltan, Wael Hamza, Yannick Versley, and Markus Boese. 2022b. Linguist: Language model instruction tuning to generate annotated utterances for intent classification and slot tagging. In *COLING 2022*.

Hsuan Su, Ting-Yao Hu, Hema Swetha Koppula, Raviteja Vemulapalli, Jen-Hao Rick Chang, Karren Yang, Gautam Varma Mantena, and Oncel Tuzel. 2023. Corpus synthesis for zero-shot ASR domain adaptation using large language models. *arXiv preprint arXiv:2309.10707*.

Peggy Tang, Kun Hu, Lei Zhang, Jiebo Luo, and Zhiyong Wang. 2023. TLDW: Extreme multimodal summarisation of news videos. *IEEE Transactions on Circuits and Systems for Video Technology*.

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.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTSCORE: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038.

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

A.1 Ethical Considerations

We rely on a multimodal large language model to create a summarization dataset that captures information from two modalities: text and images, using a self-refinement approach. While we develop a critic model trained with diverse automatic metrics and human evaluations to filter low-quality summaries for fine-tuning, we acknowledge the risk that the MLLM may generate misleading or harmful content.

We hire workers from Amazon Mechanical Turk to obtain annotations to train the critic model and compare model performance. We offer fair and timely compensation for their work. We promptly approve annotators' work and provide bonuses when applicable. We prioritize their privacy and confidentiality throughout the process.

We use all the data and resources as per their license terms. We check the distribution of categories our dataset. However, we do not analyze the dataset for potential biases.

A.2 Reproducibility Details

We generate summaries as described in Section 3.3. Step 1: We prompt the LLaVA-v1.6-Mistral-7B with the following prompt to generate a multimodal summary given a piece of text corresponding to a Wikipedia section and its related image as input: `Combine the following text with the image content and summarize coherently including content from both the text and the image and compress to a minimal length of less than three sentences such that it captures most salient information from both modalities`. We load the model in full precision to generate the summaries.

A.3 Fine-Tuning Details

We fine-tune LLaVA for 10 epochs in our self-refinement experiment which is determined by the automatic metrics on validation split of REFINESUMM. We fine-tune the multimodal projector weights of the model. In the second iteration in Sec. 5.2, we fine-tune the LLaVA model from scratch instead of the checkpoint from the first iteration. The data used to fine-tune is a combination of zero-shot summaries and summaries generated from first iteration checkpoint. We used 40 GPU hours for these fine-tuning experiments.

A.4 Confounding Nature of Automatic Metrics

Please note that the automatic evaluation metrics are designed for unimodal summaries. So we do expect a trade-off between the textual and visual metrics. Our goal in this work is to achieve a good balance of both modalities in the generated summaries while improving the correctness of the summaries.

A.5 Critic Model Details

We train an MLP classifier as the critic model with UniEval: coherence, consistency, fluency, relevance, overall, CLIPScore, BERTScore, MINT score, FactCC, Inv-TIFA as features and human rating as ground truth values.

Evaluation dimension	Precision
correct article	86.7
correct image	89.4
informative for article	87.8
informative for image	91.4

Table 6: Precision of the critic model for the chosen threshold for each of the human evaluation dimensions.

A.6 Model and Training

The terms article and text have been interchangeably in this paper and also terms LLaVA-v1.6-Mistral-7B and LLaVA. We use the LLaVA-v1.6-Mistral-7B model not the LLaVA-13B due to computational limitations. We would also like to point out that computing Inv-TIFA score for summaries is computationally expensive. It takes 15 hours to call the Claude 3 Sonnet model on AWS Bedrock, and 8 GPU hours to compute Inv-TIFA.

Note that in this work, we use the chat-based model (Liu et al., 2023b) to generate the summaries that are used for finetuning and we finetune the base model LLaVA-v1.6-Mistral-7B (Liu et al., 2024).

In Section 5.2, during the second iteration of refinement, we observe that the number of summaries passing through our screening filter increases which indicates that the self-refined summaries after iteration 1 have higher quality compared to zero-shot summaries. In the first iteration, 12k pass the filter, In the second iteration of fine-tuning, we add 12k more summaries from iteration 1 summaries that pass the filter.

Inv-TIFA: TIFA was originally designed for evaluating faithfulness of the generated image with respect to text. However, in this work, we employ it for computing faithfulness of multimodal summary (text) with respect to the image which is why we call it Inv-TIFA. To generate questions about the summary, we use Claude 3 Sonnet for question and answer generation (Anthropic, 2024). We use the BLIP-base (Li et al., 2022) model 'in the pipeline to check the correctness of the generated text with respect to the image using the questions.

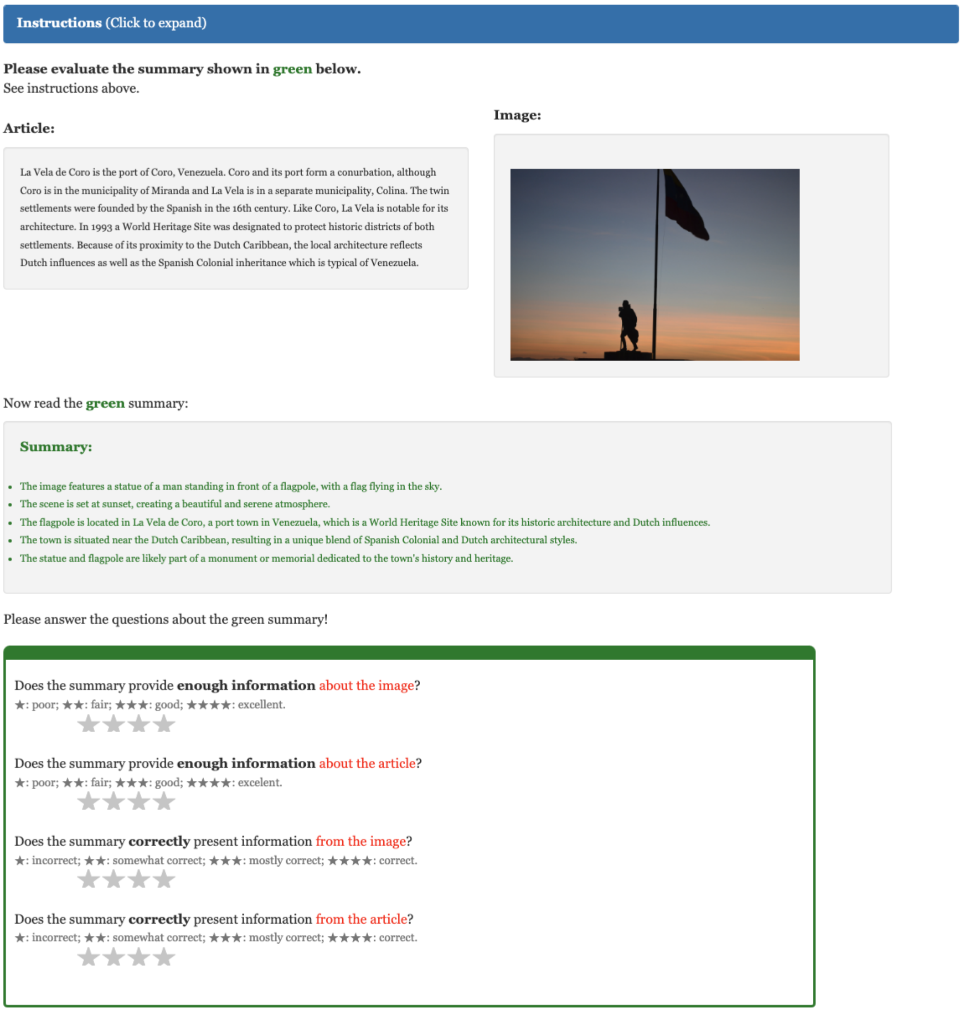

Figure 6: Mechanical Turk evaluation for informativeness and correctness of the summary with respect to the input document and image.

Modality	Metric	Zero-shot	Iter 1	Iter 2
	BERTScore	0.894	0.879	0.892
Textual	Coherence	0.949	0.879	0.903
	Consistency	0.880	0.842	0.868
	Fluency	0.944	0.942	0.926
	Relevance	0.945	0.871	0.898
	Overall	0.929	0.884	0.899
	MINT	0.597	**0.631**	0.552
	FactCC	0.612	**0.636**	0.643
Visual	CLIPScore	0.284	**0.287**	0.265
	Inv-TIFA	0.610	**0.682**	0.632
Multimodal	CLIPBERTScore	0.589	0.583	0.578

Table 7: Automatic evaluation of summaries generated zero-shot manner and each round of fine-tuning.

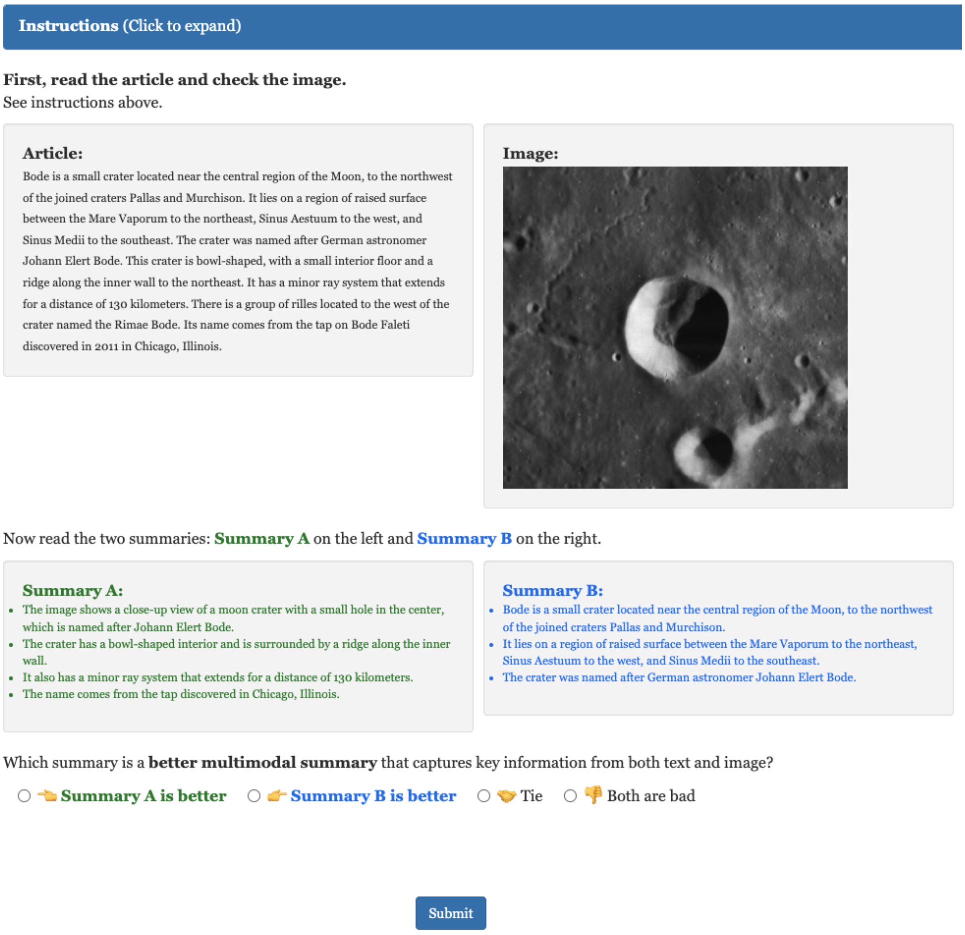

Figure 7: Mechanical Turk Evaluation for pair-wise preference annotation of multimodal summaries.

Correct Image						
threshold	precision_0	precision_1	recall_0	recall_1	fscore_0	fscore_1
0.10	0.94	0.70	0.09	1.00	0.16	0.82
0.20	0.95	0.74	0.27	0.99	0.42	0.85
0.30	0.89	0.78	0.40	0.98	0.55	0.87
0.40	0.85	0.82	0.54	0.96	0.66	0.88
0.50	0.82	0.84	0.61	0.94	0.70	0.89
0.60	0.78	0.86	0.68	0.91	0.72	0.88
0.70	0.70	0.86	0.70	0.86	0.70	0.86
0.80	0.63	0.88	0.76	0.79	0.69	0.83
0.90	0.38	0.89	0.93	0.29	0.54	0.44

Table 8: Choosing threshold such that the precision score of class 1 is greater than 0.89.