MMCIG: Multimodal Cover Image Generation for Text-only Documents and Its Dataset Construction via Pseudo-labeling

Hyeyeon Kim[1†], **Sungwoo Han**[1†], ***Jingun Kwon**[1],
Hidetaka Kamigaito[2], **and Manabu Okumura**[3]

[1]Chungnam National University
[2]Nara Institute of Science and Technology (NAIST), [3]Institute of Science Tokyo
{hyk22,77sungwhan}@o.cnu.ac.kr
jingun.kwon@cnu.ac.kr
kamigaito.h@is.naist.jp
oku@pi.titech.ac.jp

Abstract

In this study, we introduce a novel cover image generation task that produces both a concise summary and a visually corresponding image from a text-only document. Because no existing datasets are available for this task, we propose a multimodal pseudo-labeling method to construct high-quality datasets at low cost. We first collect documents with summaries, multiple images, and captions, and then exclude factually inconsistent instances. Our approach selects one image from multiple images accompanying each document. Using the gold summary, we independently rank both the images and their captions. Then, we annotate a pseudo-label for an image when both the image and its corresponding caption are ranked first in their respective rankings. Finally, we remove documents that contain direct image references within texts. Experimental results demonstrate that the proposed multimodal pseudo-labeling method constructs more precise datasets and generates higher quality images than text- and image-only pseudo-labeling methods, which consider captions and images separately.

Keywords: Multimedia Document Processing, (Semi-)Automatic Generation of Training Data, Summarization

1. Introduction

Text summarization generates concise summaries by preserving essential information from documents. Given the increasing amount of multimedia content on the web, multimodal summarization (MMS), which produces a textual summary and selects a relevant image from a document, has attracted significant attention (Li et al., 2018; Palaskar et al., 2019; Liu et al., 2020; Jangra et al., 2020; Messaoud et al., 2021; Zhuang et al., 2024).

However, existing MMS methods typically require inputs composed of both text and multiple images, making them not directly applicable to text-only scenarios commonly encountered in content creation and media. Additionally, pre-trained image generation models often struggle to produce images closely aligned with textual inputs, necessitating further fine-tuning to improve performance (Lee et al., 2023; Li et al., 2024). Moreover, creating large-scale annotated datasets of document-summary-image pairs for supervised training remains costly and challenging (Zhu et al., 2018, 2020; Jiang et al., 2023; Qiu et al., 2024).

To overcome these limitations, we propose a multimodal cover image generation (MMCIG) task that first generates concise summaries and then produces visually aligned images from text-only documents. Table 1 shows the differences between previous MMS and our MMCIG. MMCIG is directly motivated by practical needs such as thumbnail generation for news articles, where creating representative images from summaries can improve content discovery and user engagement (Zhu et al., 2018, 2020).

Aspect	MMS	MMCIG
Input at inference	Document text + ***multiple*** *images*	Document text *only*
Primary outputs	summary + ***selected*** *image*	summary + ***generated*** *image*
Need images at inference?	Yes	No
Supervision to train	(doc, multi-image, summary, gold image)	(doc, summary, gold image)

Table 1: Comparison of MMS and MMCIG. MMS assumes the availability of images during both training and inference and performs image selection, whereas MMCIG assumes text-only inputs and generates a cover image.

To support this, we introduce a multimodal pseudo-labeling method, which is the first systematic approach to constructing high-quality training

† Equal contribution
* Corresponding author

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9150–9161
11–16 May 2026.

datasets at low cost: (1) Collect documents with multiple images, their captions, and summaries from the *DailyMail* website.[1] (2) Filter documents for factual consistency. (3) Independently rank images and captions by relevance to the gold summaries. (4) Annotate an image with a multimodal pseudo-label when both the image and its corresponding caption are ranked first in their respective rankings, ensuring consistency between the textual and visual content. (5) Remove documents that explicitly reference images in their text.

We compare our multimodal pseudo-labeling method for constructing training datasets with text-only and image-only pseudo-labeling methods. In the text-only pseudo-labeling method, we rely solely on caption rankings to annotate a pseudo-label for an image. For the image-only pseudo-labeling method, we use only image rankings.

Experimental results demonstrate that our multimodal pseudo-labeling method constructs more precise datasets than the text-only and image-only methods. Furthermore, models fine-tuned on our dataset achieve improved performance in image generation. Human evaluation of both the constructed dataset and model-generated outputs confirms that our multimodal pseudo-labeling method effectively constructs precise datasets at low cost, enabling trained models to generate images closely aligned with summaries. We release our code and data at: https://github.com/HyeyeeonKim/MMCIG.

Our contribution can be summarized as follows:

- We introduce MMCIG, a practical, text-only input task that couples document summarization with image generation rather than image selection, bridging a key gap left by prior MMS formulations.
- We present a multimodal pseudo-labeling method that jointly leverages image-summary and caption-summary agreement, preceded by factuality filtering and followed by explicit image-reference filtering to reduce leakage and noise.
- We provide three datasets, $\text{MMCIG}_{\text{Text}}$ (Caption-only ranking), $\text{MMCIG}_{\text{Image}}$ (Image-only ranking), and $\text{MMCIG}_{\text{Multi}}$ (Multimodal ranking).
- Our multimodal pseudo-labeling produces more precise datasets than text-only or image-only alternatives, and models fine-tuned on $\text{MMCIG}_{\text{Multi}}$ generate higher quality, better-aligned images.

[1] https://www.dailymail.co.uk

2. Related Work

Text summarization can be categorized into two types: extractive and abstractive. Extractive summarization selects salient sentences from a given document (Cheng and Lapata, 2016; Nallapati et al., 2017; Zhou et al., 2018; Liu and Lapata, 2019), whereas abstractive summarization produces novel words and sentences (Liu and Liu, 2021; Dou et al., 2021; Liu et al., 2022; Goyal et al., 2022). Providing a summary with its corresponding image can further improve user-friendliness, as judged by human evaluation (Zhu et al., 2018, 2020); thus, significant attention has been paid to multimodal summarization. This task processes inputs from more than one modality and integrates information across different modalities to generate outputs (UzZaman et al., 2011; Bian et al., 2013; Wang et al., 2016; Li et al., 2017; Sanabria et al., 2018; Zhang et al., 2024).

Recently, Zhu et al. (2018) created the first large corpus for the multimodal summarization task (MMS). The corpus consists of two input modalities, text and multiple images, and provides a concise summary with its corresponding image. The integration of image-caption information has been proposed to better combine textual and visual features during training and inference (Zhu et al., 2020; Jiang et al., 2023). To further improve performance on the MMS task, multitask training methods (Mukherjee et al., 2022; Zhang et al., 2022b) and graph networks with hierarchical fusion frameworks for learning intra- and inter-modal correlations (Zhang et al., 2022a) have been investigated. Zhang et al. (2024) extracted entities from the text using an external knowledge graph to guide a model for image selection.

Despite the success of previous MMS methods, such methods require both text and multiple images as input. In practical scenarios where only text inputs are available, for example, in thumbnail generation for news articles, MMS cannot produce user-friendly pictorial outputs. To address this issue, we propose MMCIG, which learns to produce pictorial summaries from text-only inputs. For this novel task, we also propose a multimodal pseudo-labeling method to construct a high-quality training dataset.

3. Multimodal Pseudo-labeling

The MMCIG task applies to real-world scenarios in which only text inputs are available for user-friendly content that requires both text and images as summaries (Zhu et al., 2018, 2020). Due to the absence of suitable datasets, we propose a multimodal pseudo-labeling method to efficiently construct high-quality training datasets.

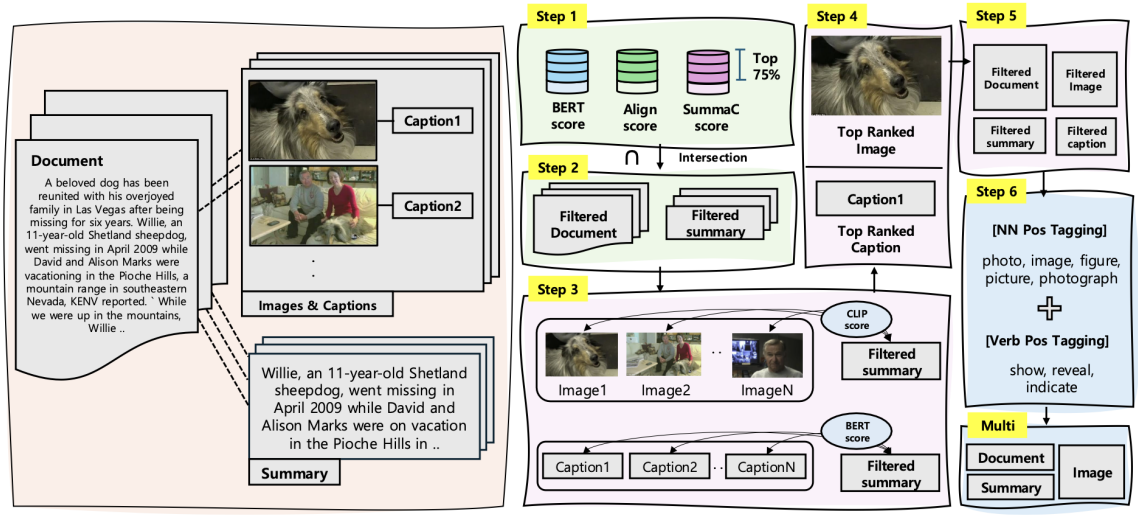

Figure 1: Overview of the MMCIG dataset construction pipeline.

Overview of Dataset Construction Pipeline. Figure 1 shows an overview of the dataset construction pipeline. We first collect a large-scale multimodal dataset from the *DailyMail* website, containing documents with multiple images, including their captions and summaries. Then, we filter out factually inconsistent instances. We independently rank both images and their captions using gold summaries, and annotate a multimodal pseudo-label for an image when both the image and its corresponding caption are ranked first in their respective rankings. Finally, we remove instances that contain direct image references within the document.

Filtering for Factual Consistency. The *DailyMail* dataset suffers from factual inconsistencies between documents and their summaries; thus, we first filter these pairs using factuality models (Guo et al., 2022): $\mathrm{BERTScore}_{\mathrm{Art}}$, AlignScore, and SummaCscore. Each model evaluates the consistency differently. $\mathrm{BERTScore}_{\mathrm{Art}}$ computes token-level similarity (Zhang* et al., 2020), AlignScore measures chunk-sentence alignment (Zha et al., 2023), and SummaCscore considers entailment scores (Laban et al., 2022). We independently remove the lowest-scoring 25% of document-summary pairs for each model and retain only the document-summary pairs that are kept across all models to ensure factuality (Guo et al., 2022). For factuality scoring, we directly compare the gold summary with the corresponding document.

Ranking Images and Captions. We independently rank the images and their captions from each document using the gold summaries, assuming captions typically provide descriptive information about the images (Jiang et al., 2023). Specifically, we rank images and captions by computing cosine similarity with summaries using CLIP (Radford et al., 2021) and BERTScore (Zhang* et al., 2020), respectively.

Algorithm 1 Filtering Algorithm.

Require:
Documents $\mathrm{Docs} = \{D_1, D_2, \ldots, D_n\}$
$\mathrm{NLTK}_{\mathrm{ssplit}}$ = sentence splitter
$\mathrm{NLTK}_{\mathrm{tagger}}$ = part-of-speech tagger
Initialize the filtered documents, F = [ ]
Initialize the $\mathrm{NN}_{\mathrm{word}}$ and $\mathrm{VB}_{\mathrm{word}}$ lists
for $i = 1$ to n **do**
  $S \leftarrow \mathrm{NLTK}_{\mathrm{ssplit}}(\mathrm{D_i})$
  bad_found $\leftarrow$ False
  for $j = 1$ to $|S|$ **do**
    $\mathrm{Tag} \leftarrow \mathrm{NLTK}_{\mathrm{tagger}}(\mathrm{S_j})$
    if $\mathrm{Tag}_{\mathrm{nn}}$ in $\mathrm{NN}_{\mathrm{word}}$ **and** $\mathrm{Tag}_{\mathrm{vb}}$ in $\mathrm{VB}_{\mathrm{word}}$ **then**
      bad_found $\leftarrow$ True
      break
    end if
  end for
  if not bad_found **then**
    $\mathrm{F.append}(\mathrm{D_i})$
  end if
end for
return F

Annotating Images with Multimodal Consistency. By independently ranking the images and captions, we obtain two separate rankings. An image is assigned a multimodal pseudo-label when both the image and its corresponding caption are ranked first in their respective rankings. This method ensures that the pseudo-labeled image is visually relevant to the summary and textually aligned through its caption.

Filtering for Direct Image Reference. Because the *DailyMail* dataset often contains direct references to image information in the documents (Hermann et al., 2015), we filter out such documents. Algorithm 1 describes how to filter such samples. We first split each document into sentences with

	MMCIG$_{\text{Text}}$	MMCIG$_{\text{Image}}$	MMCIG$_{\text{Multi}}$
Train	140,212 (555.7/55.2)	140,212 (555.7/55.2)	48,866 (504.9/55.5)
Valid	4,911 (584.6/55.1)	4,911 (584.6/55.1)	1,662 (544.7/55.8)
Test	4,968 (566.8/58.2)	4,968 (566.8/58.2)	1,774 (506.0/57.2)

Table 2: Statistics of MMCIG datasets. The numbers x/y in parentheses indicate the average document and summary lengths based on words, respectively.

	Original	After Factual Consistency	After Multimodal Consistency	After POS tagging
Train	293,966	140,212	50,496	48,866
Valid	10,353	4,911	1,726	1,662
Test	10,262	4,968	1,832	1,774

Table 3: Dataset statistics after each filtering step.

NLTK, and then tag a POS for each word.[2] Next, we build candidate lists for singular nouns (NN$_{\text{word}}$: "photo," "image," "figure," "picture," "photograph") and base-form verbs (VB$_{\text{word}}$: "show," "reveal," "indicate"). We remove documents containing sentences with both a tagged noun and a tagged verb.

MMCIG Dataset Statistics. We create three versions: MMCIG$_{\text{Text}}$ (selecting images via caption ranking only), MMCIG$_{\text{Image}}$ (via image ranking only), and MMCIG$_{\text{Multi}}$ (via both image and caption rankings). Tables 2 and 3 show the statistics of the MMCIG datasets.

Starting from the original dataset, containing 293,966 training, 10,353 validation, and 10,262 test samples, the dataset is successively filtered through stages of factual consistency, multimodal consistency, and POS tagging. After applying factual consistency filtering, the size of the dataset reduces significantly to 140,212 training, 4,911 validation, and 4,968 test instances. Subsequent filtering based on multimodal consistency further reduces the training set to 50,496, validation set to 1,726, and test set to 1,832 samples. Finally, after POS tag filtering, the dataset comprises 48,866 training, 1,662 validation, and 1,774 test samples.

For MMCIG$_{\text{Image}}$ and MMCIG$_{\text{Text}}$, we utilize the dataset obtained after applying factual consistency filtering. For MMCIG$_{\text{Multi}}$, we utilize the dataset obtained after POS tag filtering. While both MMCIG$_{\text{Text}}$ and MMCIG$_{\text{Image}}$ share identical dataset sizes due to their reliance on a single modality for labeling, MMCIG$_{\text{Multi}}$ contains a smaller number of instances.

MMCIG Dataset Evaluation. Training and validation datasets are constructed from our collected data. To evaluate the pseudo-labeling method itself, we use the MSMO test dataset (Zhu et al., 2018) since it includes multiple human-annotated gold images for each instance, where each instance consists of a document with multiple images along with its summary. Table 4 shows that MMCIG$_{\text{Multi}}$ consistently outperforms both MMCIG$_{\text{Text}}$ and MMCIG$_{\text{Image}}$ by accurately aligning summaries with their corresponding images, even when only one gold reference image is provided among multiple images in a document. Thus, our method effectively constructs high-precision datasets with closely aligned document-summary-image pairs.

[2] https://www.nltk.org/

Dataset	1	2	3	4	5	6	Avg/Total
MMCIG$_{\text{Text}}$	70.0 (298)	83.7 (788)	84.3 (779)	86.7 (670)	69.6 (1,323)	50 (3)	77.7 (3,861)
MMCIG$_{\text{Image}}$	72.3 (306)	84.9 (799)	86.2 (796)	86.9 (672)	69.5 (1,322)	66.7 (4)	78.5 (3,899)
MMCIG$_{\text{Multi}}$	**86.2** (250)	**88.2** (412)	**90.0** (316)	**88.8** (207)	**78.0** (337)	**100** (1)	**85.9** (1,523)

Table 4: Accuracy of MMCIG$_{\text{Text}}$, MMCIG$_{\text{Image}}$, and MMCIG$_{\text{Multi}}$ pseudo-labeling methods evaluated on the MSMO test dataset. Columns 1-6 represent the number of gold reference images in the documents. The numbers in parentheses represent the number of correctly annotated samples in the human-annotated MSMO test dataset.

Dataset	Alignment	Win
Random	3.60	20
MMCIG$_{\text{Multi}}$	**3.85**†	**66**

Table 5: Human evaluation results. "Win" denotes the count of pairwise higher scores. † indicates the improvement is significant (p<0.05) compared with the underlined score using paired-bootstrap-resampling with 100,000 random samples (Koehn, 2004).

We also conducted human evaluation. We sampled 100 images in the MMCIG$_{\text{Multi}}$ test dataset and created another dataset (**Random**) by randomly selecting one gold image from multiple gold images in the MSMO test dataset for the corresponding documents. We used Amazon Mechanical Turk with 80 annotators (US high school or bachelor's degree), who rated the image-summary alignment (1 to 5, 5 is the best). Table 5 shows that MMCIG$_{\text{Multi}}$ significantly outperformed **Random**, confirming our method effectively annotates images closely aligned with summaries.

4. Experiments

4.1. Experimental Settings

Datasets and Implementation Details. We used the MMCIG datasets constructed in Section 3 for the cover image generation task. Note that we used only the 1,774 test samples from MMCIG$_{\text{Multi}}$ for evaluation in all experiments to ensure data quality and consistency. For

Hyper-parameters	
Seed	42
Number of training epochs	20
Early stopping	3
Batch size	20
Image resolution	768
Optimizer	AdamW
Learning rate	3e-7
Learning rate scheduler	constant

Table 6: Hyper-parameters for open-source image generation models. DALL-E-3 was used via API without fine-tuning.

Hyper-parameters	
Seed	42
Number of training epochs	20
Early stopping	3
Batch size	8
Optimizer	AdamW
Learning rate	1e-4
Lora rank	8
Lora alpha	16
Lora dropout	0.1
Target modules	query, key, value, and output

Table 7: Hyper-parameters for text generation models.

image generation, we employed the following models: `DALL-E-3` (OpenAI, 2023), `stable-diffusion-2-1` (Rombach et al., 2022), and `dreamlike-photoreal-2.0` (Dreamlike, 2023). We used 30 inference steps with a guidance scale of 7.5 for both Diffusion-2.1 and Dreamlike.

To summarize documents, we employed the following LLMs: `Llama-3.2-3B-Instruct` (Meta, 2024) and `Qwen2.5-3B-Instruct` (Qwen et al., 2025). We fine-tuned the models on $\text{MMCIG}_{\text{Multi}}$ and used greedy decoding. Tables 6 and 7 show the hyper-parameters for fine-tuning image generation and summary generation, respectively.

While we fine-tuned parameters for open-source image generation models, we incorporated a parameter-efficient fine-tuning method (Mangrulkar et al., 2022) for summarization, specifically low-rank adapters, which combine trainable low-rank matrices with the frozen weights in transformer layers (Hu et al., 2022).

Evaluation Metrics. To assess the generated images, we considered traditional evaluation metrics, including the Fréchet Inception Distance (FID), which evaluates the distance between the probability distributions of gold and generated images (Heusel et al., 2018), and the Inception Score

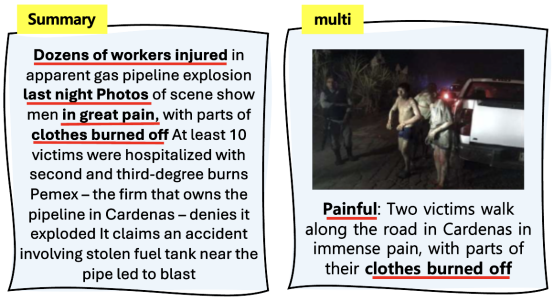

Figure 2: Example of a pseudo-labeled $\text{MMCIG}_{\text{Multi}}$ instance with its caption.

(IS), which evaluates the diversity and semantic meaningfulness of generated images (Salimans et al., 2016). In addition, we used CLIPScore to assess how well the generated images align with both the generated summaries (Txt-Img) and target images (Img-Img) (Hessel et al., 2021). Furthermore, we employed CLIP Image Quality Assessment (IQA) to measure the visual quality of images (Wang et al., 2023) and the BLIP score to evaluate how effectively the generated images align with the generated summaries (Li et al., 2022).

For summarization, we used ROUGE-1 (R-1), -2 (R-2), and -L (R-L) (Lin, 2004) to assess summarization performance. We also considered BERTScore (BS) to assess the contextual similarity between generated and gold summaries (Zhang et al., 2020).

4.2. Results

We first evaluate images generated using gold summaries and then evaluate images from generated summaries. Table 8 shows the results. While DALL-E-3 suffered difficulties to generate relevant images for summaries, We observed performance gains in BLIP, CLIP, and IQA scores with fine-tuned models compared to pre-trained models. Furthermore, $\text{MMCIG}_{\text{Multi}}$ achieved scores that were better than or comparable to $\text{MMCIG}_{\text{Text}}$ and $\text{MMCIG}_{\text{Image}}$, demonstrating the importance of constructing a high-quality dataset. While $\text{MMCIG}_{\text{Text}}$ and $\text{MMCIG}_{\text{Image}}$ are larger and potentially more diverse, results indicate that cross-modal alignment yields better performance even with fewer samples.

Fine-tuning first on $\text{MMCIG}_{\text{Image}}$ and subsequently on $\text{MMCIG}_{\text{Multi}}$ further improved performance. However, IS and FID scores exhibit inconsistencies due to their limited capability in accurately evaluating images produced by recent generative models (Jayasumana et al., 2024).

Table 9 shows the results for summarization models fine-tuned on datasets derived from $\text{MMCIG}_{\text{Multi}}$. LoRA fine-tuning yielded consistent performance gains because it enables task-

Summary Gen.	Image Gen.	Setting	MMCIG	BLIPScore (↑)	CLIPScore (↑)		IQA (↑)	IS (↑)	FID (↓)
					Txt-Img	Img-Img			
Gold	DALL-E-3	Pre-trained	-	20.0	26.4	52.1	0.84	9.3	84.3
	Diffusion-2.1	Pre-trained	-	28.8	31.5	64.2	0.97	14.8	**51.6**
		Fine-tuned	Text	29.3	31.2	65.1	0.98	**15.9**	61.1
			Image	29.8	31.7	66.0	0.98	15.8	56.8
			Multi	30.0†	32.0†	67.1†	0.98	15.1	54.3
			Image → Multi	**30.1**†	**32.1**†	**67.2**†	**0.99**	15.1	53.4
	Dreamlike	Pre-trained	-	27.8	31.5	65.2	0.95	13.5	54.9
		Fine-tuned	Text	28.8	30.9	64.9	**0.98**	15.5	58.6
			Image	29.1	31.5	65.7	**0.98**	**15.6**	53.7
			Multi	**29.2**	31.4	65.7	**0.98**†	15.4	**53.4**
			Image → Multi	**29.2**	**31.6**	**66.3**†	**0.98**	**15.6**	53.5
Llama-3.2-3B -Instruct	DALL-E-3	Pre-trained	-	20.5	26.3	51.3	0.88	9.7	79.9
	Diffusion-2.1	Pre-trained	-	28.5	31.7	63.9	0.97	15.6	**50.6**
		Fine-tuned	Text	29.1	31.2	64.4	0.98	15.9	62.7
			Image	29.8	31.9	65.4	0.98	**16.9**	55.7
			Multi	29.9	32.1†	66.4†	0.98	15.6	53.8
			Image → Multi	**30.1**†	**32.2**†	**66.8**†	**0.99**†	15.6	53.5
	Dreamlike	Pre-trained	-	28.0	30.2	63.9	0.95	13.6	52.9
		Fine-tuned	Text	28.6	31.0	64.5	**0.98**	15.6	58.1
			Image	29.1	31.5	65.5	**0.98**	**16.4**	54.9
			Multi	29.0	**31.6**†	64.5	**0.98**	14.5	**50.2**
			Image → Multi	**29.2**†	31.5	**65.7**	**0.98**†	15.9	53.9
Qwen2.5-3B -Instruct	DALL-E-3	Pre-trained	-	20.8	26.63	51.6	0.88	9.91	79.81
	Diffusion-2.1	Pre-trained	-	28.6	31.8	63.4	0.96	14.5	**52.4**
		Fine-tuned	Text	29.1	31.3	64.3	**0.98**	15.7	61.9
			Image	29.7	31.8	65.4	**0.98**	**16.4**	55.8
			Multi	29.7	32.0†	66.2†	**0.98**	16.3	54.3
			Image → Multi	**30.0**†	**32.3**†	**66.6**†	**0.98**	15.2	52.6
	Dreamlike	Pre-trained	-	27.9	30.1	63.8	0.95	13.9	53.5
		Fine-tuned	Text	28.5	31.0	64.3	**0.98**	14.9	57.7
			Image	**29.1**	31.6	65.3	**0.98**	**16.3**	54.1
			Multi	**29.1**	**31.7**	64.2	**0.98**	14.8	**50.4**
			Image → Multi	29.0	31.6	**65.4**	**0.98**	**16.3**	53.4

Table 8: Experimental results for cover image generation from gold and generated summaries using LLMs. Image → Multi indicates that the model was first fine-tuned on $\text{MMCIG}_{\text{Image}}$ and then further fine-tuned on $\text{MMCIG}_{\text{Multi}}$. The notations are the same as those in Table 5.

Model	Setting	R-1	R-2	R-L	BS
Llama-3.2-3B	Pre-trained	37.3	14.9	23.7	41.7
	Fine-tuned	**47.8**	**24.7**	**34.3**	**48.8**
Qwen2.5-3B	Pre-trained	35.0	13.0	21.9	40.3
	Fine-tuned	**45.7**	**22.8**	**32.7**	**47.4**

Table 9: Experimental results for document summarization.

Model	Setting	MMCIG	Fidelity	Alignment
DALL-E-3	Pre-trained	-	2.08	2.56
Diffusion-2.1	Pre-trained	-	2.81	2.60
	Fine-tuned	Image	**2.84**	2.78
	Fine-tuned	Image → Multi	**2.84**	**3.02**†

Table 10: Human evaluation results. The notations are the same as those in Table 5.

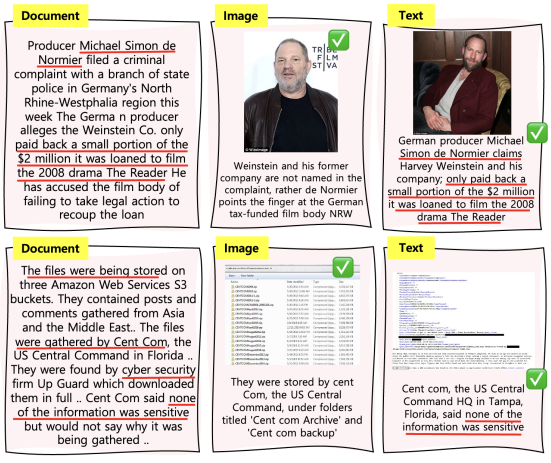

Figure 3: Two example cases showing disagreements between caption-only pseudo-labeling ($\text{MMCIG}_{\text{Text}}$) and image-only pseudo-labeling ($\text{MMCIG}_{\text{Image}}$). A check mark indicates the first image or caption based on rankings.

specific adaptation through lightweight parameter updates, consistent with prior studies on summarization (Juseon-Do et al., 2024).

4.3. Analysis

Human Evaluation. We also evaluated gold summaries and generated images by asking annotators to rate image fidelity (reality) and summary-image alignment. We sampled 100 images per model-setting combination and used Amazon Mechanical Turk with 80 annotators (US high school or bachelor's degree) to rate fidelity and alignment (1 to 5, 5 is the best). Table 10 presents the results. We

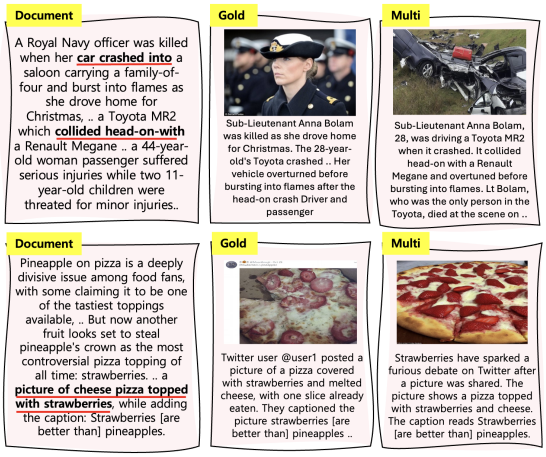

Figure 4: Example of generated images and summaries with their gold references.

observed that models fine-tuned on our MMCIG datasets significantly improved alignment between a summary and an image without compromising the fidelity. The relatively lower scores obtained by DALL-E are due to its tendency to generate images in a cartoon-like style.

Case Study for Pseudo-labeling. Figure 2 shows an example from $\text{MMCIG}_{\text{Multi}}$. The image and caption contain key information from the summary, highlighted in gray, such as *"workers injured," "last night Photos," "in great pain," and "clothes burned off."* This alignment ensures that the selected image is visually relevant to the summary and accurately represents the textual content, which enables us to construct more coherent and relevant document-image-summary pairs.

Figure 3 presents examples from $\text{MMCIG}_{\text{Text}}$ and $\text{MMCIG}_{\text{Image}}$. In $\text{MMCIG}_{\text{Image}}$, images are selected solely based on image information, whereas in $\text{MMCIG}_{\text{Text}}$, captions are considered to form document-image-summary pairs. However, ranked images and captions do not always align; thus, the constructed image-summary pairs often lack coherence and relevance (Zhu et al., 2020). In the first example, the summary relies on textual information, while the selected image does not accurately capture entities such as the person, *"Michael Simon de Normier."* In the second example, the key information *"The files were being stored"* is clearly represented in the first ranked image in $\text{MMCIG}_{\text{Image}}$, while the $\text{MMCIG}_{\text{Text}}$ image fails to capture this. This demonstrates the limitation of relying exclusively on either caption or image information.

Case Study for Generated Outputs. Figure 4 shows examples of outputs generated by Llama-3.2-3B and Diffusion-2.1 trained on $\text{MMCIG}_{\text{Multi}}$ compared to gold summaries and images from the test dataset. We observed that the model generates summaries and relevant images. In addition, providing pictorial summaries can enhance user engagement compared to textual summaries alone, which highlights the importance of the proposed task when only textual inputs are available (Zhu et al., 2018, 2020).

Figure 5 shows additional example outputs generated by the Diffusion-2.1 models from summaries generated by the fine-tuned Llama-3.2-3B including DALL-E-3. DALL-E-3 tended to produce cartoon-like images, whereas diffusion models generated more realistic images. Additionally, the model trained on $\text{MMCIG}_{\text{Multi}}$ produced images closely aligned with their corresponding summaries.

5. Discussion and Conclusion

Our primary focus is on constructing a new large-scale dataset and proposing a generation-based pipeline for cover image generation, rather than on developing or benchmarking retrieval algorithms. Retrieval-based methods typically select images from a predefined pool, while our task emphasizes generating novel cover images conditioned on textual summaries, which presents different challenges and objectives. Moreover, MMS assumes images at inference, whereas MMCIG assumes text-only inputs.

CLIP is trained through contrastive learning to align images with text and is effective for image generation (Radford et al., 2021). However, it typically uses 77 token context windows. To address this limitation, recent image generation models such as Stable Diffusion (Podell et al., 2023) and Flux (Labs, 2024) also incorporate the encoder from T5 (Raffel et al., 2023). While the fine-tuned portion can handle up to 256 tokens (and technically accepts up to 512), only the first 256 tokens are learned in a way that supports meaningful alignment, and there is no guarantee of strong alignment beyond that point (Ozaki et al., 2025). Thus, instruction-tuned text-to-image models are not suitable when conditioned on the full article. Moreover, the generated images are required to be closely aligned with summaries rather than full articles (Zhu et al., 2018, 2020).

In this paper, we proposed a novel task of MMCIG that generates textual summaries and their corresponding images from text-only documents to output user-friendly content. Due to the lack of available datasets for this task, we proposed a multimodal pseudo-labeling method to efficiently construct high-quality training datasets. Models trained on our datasets produce informative summaries accompanied by visually corresponding images, as confirmed by automatic and human evaluations.

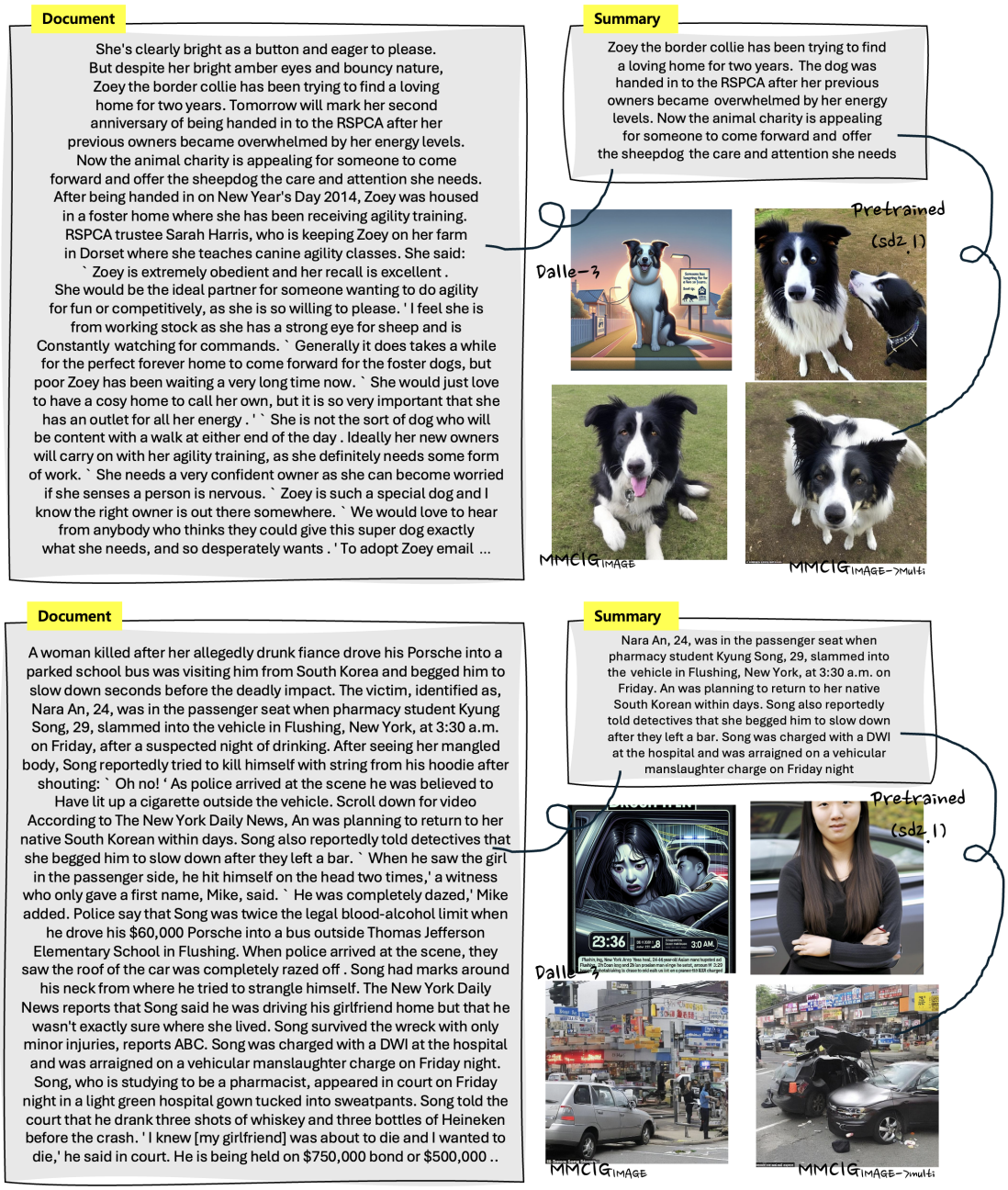

Figure 5: Another example of generated images and summaries.

Limitations

Although we proposed the MMCIG task and demonstrated the effectiveness of the proposed multimodal pseudo-labeling method, several limitations remain.

First, constructing MMCIG$_{\mathrm{Multi}}$ requires documents with multiple images and their corresponding captions. This may limit the applicability to other domains or languages that lack such rich multimodal datasets. In addition, we primarily considered the dataset from *DailyMail*, which may cause biases related to content style or cultural context due to the nature of this specific domain. However, our approach is the first systematic method for constructing pseudo-labels based on a multimodal approach. In the future, we plan to construct multilingual datasets for different domains for the proposed task.

Second, our current POS- and rule-based filter may miss synonyms, plural forms, and contextual cues, leading to false positives and negatives. We plan to enhance it by considering a hybrid approach that combines regex and semantic search.

Third, when generating images for named entities, such as specific people mentioned in the generated summaries, our image generation module struggles to accurately generate images corre-

sponding to these named entities. This may be due to challenges in learning the visual representations for less common entities. In the future, we plan to incorporate external resources to improve image generation for named entities.

Ethics Statement

This section considers the potential ethical issues associated with our model. We proposed MMCIG for the cover image generation task, which is trained on $\text{MMCIG}_{\text{Multi}}$. $\text{MMCIG}_{\text{Multi}}$ was constructed from the *DailyMail* dataset, which is a publicly available summarization dataset. Therefore, MMCIG might produce incorrect summaries and images that reflect biases present in the dataset. To mitigate these issues, we cleaned the dataset using factuality models to reduce incorrect or misleading content, since our model generates images based on textual summaries. However, this may not remove all biases present in the dataset. In the future, we plan to consider bias detection methods to better construct $\text{MMCIG}_{\text{Multi}}$.

6. Bibliographical References

Jingwen Bian, Yang Yang, and Tat-Seng Chua. 2013. Multimedia summarization for trending topics in microblogs. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, page 1807–1812, New York, NY, USA. Association for Computing Machinery.

Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–494. Association for Computational Linguistics.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. GSum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4830–4842, Online. Association for Computational Linguistics.

Dreamlike. 2023. Dreamlike photoreal 2.0.

Tanya Goyal, Jiacheng Xu, Junyi Jessy Li, and Greg Durrett. 2022. Training dynamics for text summarization models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2061–2073, Dublin, Ireland. Association for Computational Linguistics.

Yanzhu Guo, Chloé Clavel, Moussa Kamal Eddine, and Michalis Vazirgiannis. 2022. Questioning the validity of summarization datasets and improving their factual consistency. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5716–5727, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend.

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

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2018. Gans trained by a two time-scale update rule converge to a local nash equilibrium.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Anubhav Jangra, Sriparna Saha, Adam Jatowt, and Mohammad Hasanuzzaman. 2020. Multi-modal summary generation using multi-objective optimization. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1745–1748, New York, NY, USA. Association for Computing Machinery.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. 2024. Rethinking fid: Towards a better evaluation metric for image generation.

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

Juseon-Do Juseon-Do, Hidetaka Kamigaito, Manabu Okumura, and Jingun Kwon. 2024. InstructCMP: Length control in sentence compression through instruction-based large language

models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8980–8996, Bangkok, Thailand. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Black Forest Labs. 2024. Flux. https://github.com/black-forest-labs/flux.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. 2023. Aligning text-to-image models using human feedback.

Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, and Chengqing Zong. 2018. Multi-modal sentence summarization with modality attention and image filtering. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 4152–4158. AAAI Press.

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

Jialu Li, Jaemin Cho, Yi-Lin Sung, Jaehong Yoon, and Mohit Bansal. 2024. SELMA: Learning and merging skill-specific text-to-image experts with auto-generated data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.

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

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

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740. Association for Computational Linguistics.

Yixin Liu and Pengfei Liu. 2021. SimCLS: A simple framework for contrastive learning of abstractive 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 2: Short Papers)*, pages 1065–1072, Online. Association for Computational Linguistics.

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

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

Safa Messaoud, Ismini Lourentzou, Assma Boughoula, Mona Zehni, Zhizhen Zhao, Chengxiang Zhai, and Alexander G. Schwing. 2021. Deepqamvs: Query-aware hierarchical pointer networks for multi-video summarization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 1389–1399, New York, NY, USA. Association for Computing Machinery.

Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models.

Sourajit Mukherjee, Anubhav Jangra, Sriparna Saha, and Adam Jatowt. 2022. Topic-aware multimodal summarization. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 387–398, Online only. Association for Computational Linguistics.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents.

OpenAI. 2023. Dall·e 3 is now available in chatgpt plus and enterprise.

Shintaro Ozaki, Kazuki Hayashi, Yusuke Sakai, Jingun Kwon, Hidetaka Kamigaito, Katsuhiko Hayashi, Manabu Okumura, and Taro Watanabe. 2025. Texttiger: Text-based intelligent generation with entity prompt refinement for text-to-image generation.

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.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis.

Jielin Qiu, Jiacheng Zhu, William Han, Aditesh Kumar, Karthik Mittal, Claire Jin, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Ding Zhao, Bo Li, and Lijuan Wang. 2024. Mmsum: A dataset for multimodal summarization and thumbnail generation of videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21909–21921.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A large-scale dataset for multimodal language understanding.

Naushad UzZaman, Jeffrey P. Bigham, and James F. Allen. 2011. Multimodal summarization of complex sentences. In *Proceedings of the 16th International Conference on Intelligent User Interfaces*, IUI '11, page 43–52, New York, NY, USA. Association for Computing Machinery.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. 2023. Exploring clip for assessing the look and feel of images. In *AAAI*.

William Yang Wang, Yashar Mehdad, Dragomir R. Radev, and Amanda Stent. 2016. A low-rank approximation approach to learning joint embeddings of news stories and images for timeline summarization. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 58–68, San Diego, California. Association for Computational Linguistics.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Litian Zhang, Xiaoming Zhang, and Junshu Pan. 2022a. Hierarchical cross-modality semantic correlation learning model for multimodal summarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11676–11684.

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*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Yanghai Zhang, Ye Liu, Shiwei Wu, Kai Zhang, Xukai Liu, Qi Liu, and Enhong Chen. 2024. Leveraging entity information for cross-modality correlation learning: The entity-guided multimodal summarization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9851–9862, Bangkok, Thailand. Association for Computational Linguistics.

Zhengkun Zhang, Xiaojun Meng, Yasheng Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. 2022b. Unims: A unified framework for multimodal summarization with knowledge distillation.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–663. Association for Computational Linguistics.

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

Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020. Multimodal summarization with guidance of multimodal reference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9749–9756.

Haojie Zhuang, Wei Emma Zhang, Leon Xie, Weitong Chen, Jian Yang, and Quan Sheng. 2024. Automatic, meta and human evaluation for multimodal summarization with multimodal output. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7768–7790, Mexico City, Mexico. Association for Computational Linguistics.