ModalSyncSum: Synchronizing Image and Text for Reliable Summary Generation

Xuanqi Chen[1], Ziying Rong[1], Xinfeng Liao[1], Yiqian Wu[1], Bowei Zhang[1], Pengfei Fu[1], Shengyi Jiang[1,2*]

[1]School of Information Science and Technology, Guangdong University of Foreign Studies, Guangzhou, Guangdong, China
[2]Faculty of Data Science, City University of Macau, Macao Special Administrative Region of China
chenxuanqi6@163.com, RongZiYing1016@163.com, liaoxinfeng218@163.com, 13570131695@163.com, ashen_zhang@foxmail.com, Jiangshengyi@163.com

Abstract

Multimodal summarization with multimodal output (MSMO) aims to generate coherent textual summaries while selecting the most semantically relevant images to enhance expressiveness. Despite the advancements of large multimodal models like GPT-4o, LLaMA-3, and Grok-3, these models often exhibit hallucination and weak visual-text alignment when applied to MSMO tasks. To address these challenges, we propose ModalSyncSum, a unified framework that enhances semantic consistency and visual faithfulness. It incorporates image-aware information extraction to mitigate visual-text misalignment, QA-based description verification to detect and correct hallucinated image descriptions, and named entity-guided refinement to ensure factual accuracy and entity alignment across modalities. Furthermore, we introduce a new evaluation metric M^3AS, which jointly considers image content coverage, text-image alignment, and summary consistency, filling the gap in evaluating multimodal summary quality. Experimental results show that our model outperforms prompt-based baselines across multiple datasets, achieving significant gains on ROUGE, BLEU, and BERTScore, with BLEU improving by 21.95%. In human evaluation, M^3AS exhibits stronger correlation with human judgments in consistency, image-summary relevance, and focus, surpassing existing automatic metrics.

Introduction

Multimodal summarization aims to generate outputs that combine textual summaries and image descriptions, enhancing the expressiveness of conveyed content (Zhu et al. 2018). Prior studies have shown its value in applications such as news reporting, educational videos, and medical documentation (Jangra et al. 2020).

With the rise of large-scale pretrained multimodal models such as GPT-4o, LLaMA-3, and Grok-3, text summarization performance has made substantial progress, with generated summaries achieving near-human or even superhuman quality on various abstractive summarization benchmarks (Pu, Gao, and Wan 2023). These models exhibit impressive capabilities in capturing complex linguistic structures and generating coherent and informative summaries.

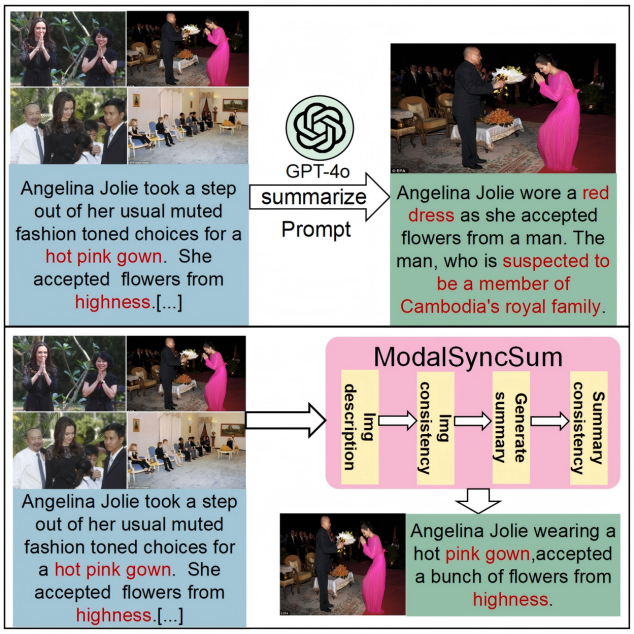

Figure 1: Comparison of summary generation quality between large multimodal models and our proposed ModalSyncSum framework, highlighting hallucination issues. Prompt-based large models often hallucinate, e.g., mistaking "hot pink" for "red" or misinterpreting the title "highness" as implying Cambodian royalty, leading to factual errors. In contrast, ModalSyncSum ensures better semantic consistency and visual-text alignment.

However, in the task of multimodal summarization with multimodal output (MSMO), these models still face significant challenges. As illustrated in Figure 1, they are prone to hallucinations, often misinterpreting visual content and producing summaries that contradict the actual image semantics. Additionally, visual-textual misalignment remains prevalent, where generated texts fail to faithfully reflect visual information, exposing limitations in current fusion and alignment strategies (Jing et al. 2023).

Moreover, existing evaluation metrics such as ROUGE and BLEU primarily assess lexical overlap and are insufficient to capture visual semantics embedded in multimodal

*Corresponding author.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

summaries. Although text-based consistency metrics like UniEval have been proposed for textual summarization, the multimodal summarization domain lacks effective metrics to assess consistency across modalities. This absence makes it difficult to holistically evaluate the accuracy and coherence of generated summaries from both textual and visual perspectives.

To address the challenges of semantic consistency and visual information integration in multimodal summarization, this paper proposes a novel summarization framework centered on multimodal consistency, aiming to generate summaries that accurately and coherently reflect both textual and visual content. The main contributions of this work are as follows:

- We propose a novel framework that does not require fine-tuning large multimodal models. It first generates image-related textual descriptions and then employs a vision-language pretrained model to extract image-aware textual segments. These are further validated by a visual question-answering (VQA) model to assess the consistency between image and generated text. This QA-based verification is used to guide iterative refinement of the summary, significantly improving consistency and visual grounding of the output.
- We design a comprehensive evaluation metric named M^3AS, which assesses multimodal summaries from three complementary perspectives: image information coverage, visual-text semantic alignment, and cross-modal consistency. This addresses the limitations of traditional text-only metrics (e.g., ROUGE, BLEU) by incorporating visual dimensions into the evaluation process.
- Experimental results on multiple MSMO datasets show that our framework significantly outperforms prompt-based approaches using state-of-the-art large models, with notable improvements on ROUGE, BLEU, and BERTScore. In human evaluation, M^3AS exhibits high correlation with human judgments on key aspects such as consistency, image-summary relevance, and focus, validating its effectiveness as an evaluation metric.

Related Work

Traditional Multimodal Summarization

Traditional multimodal summarization methods rely on modality alignment and fusion techniques—such as joint embedding and attention mechanisms—to integrate textual and visual information. Zhu et al. (2018) proposed a multimodal optimization framework (MOF) using ROUGE and Order-ranking to jointly optimize image and text tasks. Zhang et al. (2023) introduced CISum, a cross-modal framework that employs attention and filtering to generate textual summaries, visual descriptions, and image selections via a multi-task Transformer. Tang et al. (2024) leveraged Maximum Mean Discrepancy (MMD) with CLIP and learnable filters to enhance modality fusion via BART-based cross-modal attention. Despite their effectiveness, these methods often suffer from modality imbalance, where visual data is underutilized.

Multimodal Summarization with Large Language Models

The rise of large language models has shifted summarization from fine-tuning approaches toward zero-shot generation. Pu, Gao, and Wan (2023) demonstrated that LLMs significantly improve coherence and readability over traditional models. In multimodal contexts, the CLIPSyntel(Ghosh et al. 2024) framework combines CLIP's visual encoding with LLMs for medical summarization, enhancing modality complementarity. Wang et al. (2023) introduced cross-modal knowledge-guided models using graph neural networks to improve contextual consistency. Adrakatti (2024) addressed screen content summarization using alignment techniques for complex inputs. Tang et al. (2024) further improved cross-modal fusion via MMD-based alignment and feature selection. Nonetheless, challenges remain in adapting to cross-modal data distributions, mitigating visual noise, and ensuring semantic accuracy in generated outputs.

Summarization Evaluation

Summarization quality is often evaluated using ROUGE, BLEU, and BERTScore. While ROUGE focuses on n-gram overlap, BERTScore measures semantic similarity via contextual embeddings. InfoLM(Colombo, Clavel, and Piantanida 2022) extends this by applying information-theoretic divergence measures (e.g., KL divergence) using masked language models, offering robust and flexible semantic evaluation without requiring layer selection. MTEQA uses question-answering to assess information retention, particularly in multilingual scenarios (Krubiński et al. 2021).

In recent years, researchers have proposed M-info(Xiao et al. 2025), an evaluation metric specifically designed for multimodal summarization. M-info measures the consistency between the information distribution of the generated summary and that of the input content using Kullback–Leibler (KL) divergence. However, despite its effectiveness in evaluating information coverage, M-info falls short in detecting hallucinations in the generated summaries, limiting its reliability in assessing factual consistency.

LLM-based evaluation methods have shown superior performance in fluency and coherence. Prompt-based methods like GPTScore (Fu et al. 2024) estimate generation probability without training, while tuning-based approaches (e.g., X-EVAL(Li et al. 2024)) fine-tune open-source models for multidimensional evaluation . Despite their strength, such methods are sensitive to prompt design, computationally expensive, and currently lack the capacity to assess visual quality in multimodal summarization.

Methodology

This section presents our proposed multimodal summarization framework, **ModalSyncSum**, as illustrated in Figure 2, along with the evaluation metric **M^3AS**(Multimodal Triple-factor Assessment Score), designed to assess multimodal consistency.

Generate Image Descriptions

Given a news article $A = \{s_1, s_2, ..., s_m\}$, where s_i denotes the i-th sentence in the article, and an accompanying set of images $I = \{I_i, I_2, ...I_n\}$, where each image I_j corresponds to visual content associated with the article, the goal is to generate a coherent and semantically relevant textual description for each image based on the article content.

Each sentence s_i and image I_j is first encoded into a shared semantic space using a pre-trained CLIP model, producing embedding vectors $v_{s_i} \in R^d$ and $v_{I_j} \in R^d$, respectively. The similarity between sentence s_i and image I_j is then computed using cosine similarity:

$$sim(s_i, I_j) = \frac{\mathbf{v}_{s_i} \cdot \mathbf{v}_{I_j}}{\|\mathbf{v}_{s_i}\|, \|\mathbf{v}_{I_j}\|} \tag{1}$$

A similarity threshold τ_1 is defined to filter sentence-image pairs. A sentence s_i is selected as a reference for image I_j if and only if the cosine between s_i and I_j is greater than or equal to the threshold τ_1. The set of reference sentences for image I_j is denoted as:

$$\mathcal{S}_j = \{s_i \in A \mid sim(s_i, I_j) \geq \tau_1\}, \tag{2}$$

The selected reference sentence set S_j is then used as input to a large vision-language model (LVLM) to generate a description d_j corresponding to image I_j.

Consistency of Picture Description

To ensure the consistency of generated image descriptions, we adopt a dual-verification strategy utilizing the CLIP model and the image question-answering model BLIP.

In the initial stage, multiple candidate descriptions $D_j = \{d_1^j, d_2^j, \ldots, d_m^j\}$ are generated for each image I_j. The CLIP model is then used to compute the semantic similarity between each candidate description d_i^j and image I_j, following the same cosine similarity formulation as in Equation (1). Descriptions with similarity scores below a predefined threshold τ_2 are discarded:

$$\mathcal{D}_j^{CLIP} = \left\{ d_i^j \in D_j \;\middle|\; sim(d_i^j, I_j) \geq \tau_2 \right\}, \tag{3}$$

Next, to verify factual alignment with the image content, we construct question prompts Q_i from each retained description $d_i^j \in D_j^{CLIP}$, focusing on fine-grained visual attributes (e.g., clothing, facial expressions, actions, background objects). These questions are answered by the BLIP model based on the visual input I_j:

$$\mathcal{B}_i = BLIP(I_j, Q_i), \tag{4}$$

We then assess the consistency between the generated description and the BLIP-provided answers. Only those descriptions that are semantically consistent with the answers are retained:

$$\mathcal{D}_j^{\text{Final}} = \left\{ d_i \in \mathcal{D}_j^{\text{CLIP}} \;\middle|\; Consis(d_i, \mathcal{B}_i) = \text{True} \right\}, \tag{5}$$

Finally, the high-quality descriptions in $\mathcal{D}_j^{\text{final}}$ are aggregated by prompting the language model to merge them into a complete and coherent final description d_j^*.

Summary Generation and Consistency Verification

To ensure fluency and coherence in the generated textual summary, we first embed the image descriptions into the article based on semantic similarity. For each image description d_j^*, we compute its similarity with every sentence s_i. The sentence with the highest similarity is identified as the insertion anchor:

$$i^* = \arg\max_i \text{sim}(d_j^*, s_i) \tag{6}$$

The modified article A^{mod} is then constructed by inserting d_j^* immediately after s_{i^*}. This results in a structurally coherent article enriched with visual context. Finally, the news summary y is generated through LLM based on the content of A^{mod}.

To further ensure factual consistency between the generated summary y and the original article D, we propose a QA-based verification framework. We begin by extracting a set of named entities $\mathcal{E}$ from the summary using a named entity recognition model. Next, entity-specific questions $Q^{\mathcal{E}} = \{q_1^{\mathcal{E}}, q_2^{\mathcal{E}}, ...q_k^{\mathcal{E}}\}$ are automatically generated using prompt-based instructions.

Each question $q_i^{\mathcal{E}}$ is answered using two sources, from the summary a_i^y and from the article a_i^A. We then compare both answers for consistency:

$$Consis(q_i^{\mathcal{E}}) = \begin{cases} \text{True}, & \text{if } a_i^y \approx a_i^A \\ \text{False}, & \text{otherwise} \end{cases} \tag{7}$$

If inconsistency is detected, $\exists\, q_i^{\mathcal{E}}$; Consis$(q_i^{\mathcal{E}}) = $ False, the LLM is prompted to revise the answer or regenerate the summary:

$$y \leftarrow LLM\text{-}Revise\left(y, \left\{q_i^{\mathcal{E}}, a_i^{\mathcal{A}}\right\}\right). \tag{8}$$

This consistency-checking process is iteratively repeated until:$\forall\, q_i^{\mathcal{E}} \in \mathcal{Q}^{\mathcal{E}},\ Consis(q_i^{\mathcal{E}}) = \text{True}$.

M^{3}AS

To comprehensively evaluate the accuracy of both visual and textual information in multimodal summarization tasks, we propose a novel evaluation metric: **M^{3}AS** (Multimodal Triple-factor Assessment Score). This metric assesses the quality of a generated summary from three dimensions: (1) the semantic content of the image, (2) the semantic similarity between the image and the text, and (3) the consistency of the summary content.

Image information coverage measures the proportion of key information from the image description that appears in the generated summary:

$$\text{Img}_{\text{coverage}} = \frac{|\{k \mid k \in \mathcal{K}_{\text{img}} \wedge k \in \mathcal{K}_y\}|}{|\mathcal{K}_{\text{img}}|}, \tag{9}$$

where $\mathcal{K}_{\text{img}}$ denotes the set of key image-related information units (e.g., entities, events, relations) from the image description D_j^{Final}, and $\mathcal{K}_y$ denotes the set of key information units in the generated summary y.

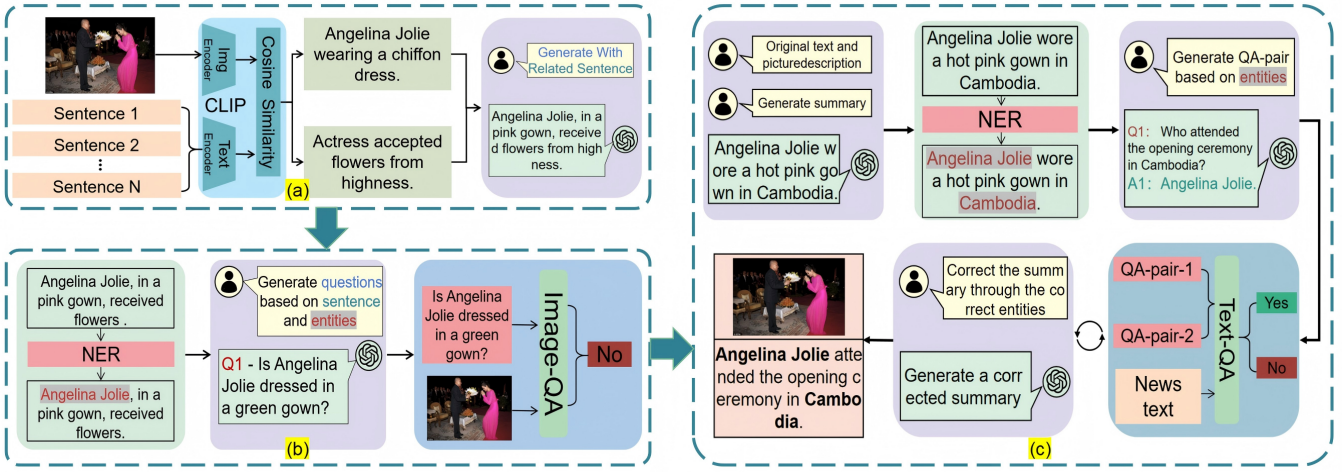

Figure 2: An overview of the ModalSyncSum model. (a) CLIP identifies image-relevant sentences and prompts the large model to generate descriptions; (b) Named entity recognition (NER) and image QA validate the description's accuracy; (c) The description is embedded into the article, guiding summary generation, which is further verified and refined using NER and text-based QA.

Image information density measures the ratio of image-related information to the total key information in the summary:

$$\text{Img}_{\text{density}} = \frac{|\{k \mid k \in \mathcal{K}_{\text{img}} \wedge k \in \mathcal{K}_y\}|}{|\mathcal{K}_y|}, \tag{10}$$

$$\text{Score}_{\text{img\_info}} = \sqrt{\text{Img}_{\text{coverage}} \cdot \text{Img}_{\text{density}}} \tag{11}$$

Image Information Score $\text{Score}_{\text{img\_infor}} \in [0, 1]$ integrates both image information coverage and density. The score is designed to be sensitive—if either metric is significantly low, the overall score will decrease sharply. A higher $\text{Score}_{\text{img\_infor}}$ indicates that the multimodal summary not only provides comprehensive coverage of image content but also maintains high compactness and relevance with respect to the image.

$$\text{Score}_{\text{img\&sum}} = \sqrt{sim(d^*, y) \cdot sim(I, y)} \tag{12}$$

$\text{Score}_{img\&sum} \in [0, 1]$ measures the semantic consistency between the image and the summary. I, y, and d^* represent images, summary, and images description, respectively. sim formula as shown in Equation (1). The score combines cosine similarity to assess how well the image description semantically aligns with the summary. Higher scores indicate better alignment between the image content and the summary.

$$\text{Score}_{\text{consist}} = \frac{\sum_{i=1}^{N} \left(r_i \cdot c_i \cdot \frac{m_i}{M} \cdot \delta\right)}{\sum_{i=1}^{N} c_i} \tag{13}$$

$\text{Score}_{\text{consist}} \in [0, 1]$ measures the consistency between a multimodal summary and the original text-image content. Let $\mathcal{E} = \{e_1, \ldots, e_N\}$ be the set of key summary points, each converted into a question-answer (QA) pair. The answer correctness $r_i \in 0, 1$ is determined using the answer matching function defined in Equation (7). Each QA pair is weighted by c_i, the frequency of e_i in the original text A and image description D^{Final}, reflecting its content importance. Cross-modal alignment is rewarded via $\frac{m_i}{M}$, where $m_i \in [0, M]$ is the number of modalities in which e_i appears, and M is the total number of modalities. δ acts as a temperature coefficient to control the strength of cross-modal preference, a higher δ increases sensitivity to multimodal alignment.

This score jointly reflects semantic accuracy, salience, and modality alignment. Higher values indicate better consistency between the summary and the original multimodal content.

$$\begin{aligned} \text{M}^3\text{AS} = \alpha \cdot \text{Score}_{\text{img\_info}} + \beta \cdot \text{Score}_{\text{img\&sum}} \\ + \gamma \cdot \text{Score}_{\text{consist}} \end{aligned} \tag{14}$$

where α, β, and γ are weights summing to 1, ensuring a balanced assessment of image contribution in terms of content coverage and consistency.

Experiment

Datasets

We conduct experiments on the following datasets:

MSMO(Zhu et al. 2018): A multimodal summarization dataset derived from the Daily Mail website, featuring manually written bullet-point summaries and associated images selected by graduate students.

M3LS(Verma et al. 2023): A multilingual multimodal summarization dataset, organized by language, containing articles with their corresponding images and metadata.

E-Liputan(Song et al. 2024): A multimodal dataset in Indonesian, designed for summary-guided generative sum-

marization, aimed at addressing the lack of Indonesian-language resources for multimodal summarization models.

Baselines

We compared our model with a wide range of state-of-the-art unimodal and multimodal summarization baselines. For multimodal baselines, we evaluate against **Vision-GPLM**(Jing et al. 2023) and **VG-GPLMs**(Yu et al. 2023), which integrate visual information into language models for enhanced generation. We also include **Va-SOGM**(Song et al. 2024), an end-to-end two-stage model with summary-guided visual-text co-attentionl. Additionally, large-scale general-purpose models like **GLM-4V**(Zeng et al. 2024), **Grok-3**(xAI Team 2024), **Gmini**(Google DeepMind 2024), **Doubao-Vision-Pro**(ByteDance-Seed Team 2024), and **GPT-4o**(OpenAI 2024) are compared for their cross-modal generation capabilities.

Implementation Details

In this experiment, we use the pretrained **CLIP**[1] model to calculate semantic similarity between images and descriptions. We apply **CLIP-ViT**[2] and **BERT-SQuAD**[3] for QA validation to check consistency between image descriptions and summaries, and **BERT-NER**[4] for named entity recognition to extract key text information.

We set two thresholds for validation: $\tau_1 = 0.3$ for image-description similarity and $\tau_2 = 0.4$ for summary-source QA consistency. For the final scoring function, we set the fusion weights as $\alpha = 0.25$, $\beta = 0.25$, and $\gamma = 0.5$. To evaluate performance, we sampled 5,000 instances from each of the three benchmark datasets, using stratified sampling for the multilingual M3LS to ensure balanced language coverage.

Evaluation Metrics

We employ both automatic and human evaluation metrics to assess the quality of the generated summaries. For automatic evaluation, we use **ROUGE** (Lin 2004), a standard summarization metric that measures n-gram overlap with a focus on content recall; **BLEU** (Papineni et al. 2002), which evaluates lexical and structural similarity based on n-gram precision; and **BERTScore** (Zhang et al. 2020), which leverages contextual embeddings from pre-trained BERT models to capture semantic similarity beyond surface-level matches. Additionally, we introduce **M^{3}AS**, a novel multimodal summarization metric proposed in this work . For human evaluation, we adapt the criteria from (Koto, Baldwin, and Lau 2022), (Fabbri et al. 2021), and (Krubiński and Pecina 2023), using four dimensions rated on a 0–4 Likert scale: **Consistency** (factual alignment with the source content), **Fluency** (grammaticality and readability), **Focus&Coverage** (inclusion of key information and overlap with reference content), and **Image-Summary Relevance** (semantic and visual alignment between images and the summary).

[1]https://huggingface.co/openai/clip-vit-large-patch14

[2]https://huggingface.co/sentence-transformers/clip-ViT-B-32-multilingual-v1

[3]https://huggingface.co/google-bert/bert-large-uncased-whole-word-masking-finetuned-squad

[4]https://huggingface.co/Davlan/bert-base-multilingual-cased-ner-hrl

Result and Analysis

Overall Performance

As shown in Table 1, although ModalSyncSum yields relatively lower R-1, R-2, and BLEU scores compared to purely text-based models and conventional multimodal summarization methods, it performs better in terms of BERTScore and M^{3}AS. This indicates that the summaries generated by ModalSyncSum diverge more from the reference summaries at the surface level but achieve higher accuracy in semantic alignment and image-text consistency. Equipped with the proposed ModalSyncSum framework, LVLMs achieve notable relative improvements across all metrics compared to direct prompting. On average, ModalSyncSum leads to +21.3% improvement in R-1, +14.7% in R-2, +15.4% in R-L, +18.1% in B-1, +25.8% in B-2, +7.8% in BERTScore, and +23.5% in M^{3}AS. These results demonstrate that ModalSyncSum effectively enhances LVLM summarization quality, achieving improved lexical and semantic alignment.

Table 2 presents the performance of our proposed multimodal summarization framework, ModalSyncSum, compared to several state-of-the-art baselines on MSMO, M3LS, and E-Liputan. ModalSyncSum achieves consistently strong results across all datasets, particularly excelling in BERTScore and M^{3}AS, indicating its advantage in semantic consistency and multimodal alignment.

In cross-lingual and multilingual settings, ModalSyncSum performs slightly lower on M3LS and E-Liputan compared to MSMO, likely due to the limitations of current multimodal pre-trained models in low-resource languages. Nevertheless, its competitive performance on BERTScore and M^{3}AS suggests strong generalization and language-agnostic multimodal understanding. Among the metrics, M^{3}AS provides a more task-aligned evaluation by emphasizing cross-modal semantic consistency and fusion quality, making it particularly suitable for assessing real-world multimodal summarization systems.

Ablation Study

As shown in Table **??**, GPT-4o (Img) performs worse than the text-only GPT-4o across all metrics, indicating that directly prompting LVLMs with images may even degrade summarization quality due to insufficient task alignment. In contrast, ModalSyncSum significantly improves performance over GPT-4o (Img), with +12.7 in R-1, +6.8 in R-L, +4.8 in B-1, +6.3 in BERTScore, and +6.5 in M^{3}AS, demonstrating the effectiveness of the proposed framework in enhancing cross-modal summarization.

We further examine the impact of removing the image and summary consistency checking modules. Ablating either component resultd in noticeable performance degradation, particularly on BERTScore and M^{3}AS, confirming their roles in improving semantic coherence and cross-modal alignment. Overall, the complete ModalSyncSum framework consistently outperforms its ablated variants across all

	Model	R-1 [%]↑	R-2 [%]↑	R-L [%]↑	B-1 [%]↑	B-2 [%]↑	BERTScore [%]↑	M^3AS [%]↑
Traditional Multi-modal Summarization	Vision-GPLM	36.92±0.3	23.10±1.1	31.09±1.2	17.38±1.8	14.75±0.3	34.18±1.4	18.58±0.6
	VG-GPLMs	40.16±1.1	25.47±0.3	32.09±1.5	20.69±0.8	16.01±0.5	48.94±2.3	23.19 ±0.6
	Va-SOGM	**41.94±1.8**	**28.15±0.6**	32.60±0.5	**22.32±0.6**	**18.13±0.4**	45.34±1.7	23.43±0.9
LVLM + Prompt	GPT-4o	27.22±1.3	18.54±1.2	25.83±1.3	17.46±1.0	10.51±0.4	54.37±2.7	28.93±1.0
	GLM-4v	29.12±0.8	19.49±0.4	26.16±0.2	18.70±1.5	11.72±0.1	56.51±2.2	26.18±1.2
	Grok-3	28.90±1.4	19.55±0.7	25.84±1.6	16.91±0.4	11.67±0.6	52.03±1.8	27.42±0.6
	Gmini	24.06±1.7	17.43±0.3	23.24±0.6	15.44±0.3	9.36±0.2	54.31±1.5	25.86±0.3
	Doubao	26.32±0.2	18.93±1.3	24.57±0.4	18.53±0.3	10.48±0.4	51.48±2.3	26.20±1.4
ModalSyncSum	GPT-4o	31.31±0.8	19.06±1.1	25.44±0.1	18.15±0.8	16.96±0.1	59.48±2.9	32.93±0.5
	GLM-4v	39.95±1.0	25.74±0.9	**32.64±0.4**	22.25±1.4	18.03±0.6	**60.71±1.4**	**35.46±0.7**
	Grok-3	37.23±1.5	24.06±0.1	31.86±0.4	21.45±0.1	17.93±0.4	54.52±2.0	29.43±0.2
	Gmini	35.10±1.3	21.39±0.1	28.24±0.6	21.44±0.8	17.36±0.2	57.31±1.6	32.74±0.4
	Doubao	36.49±0.4	22.18±0.7	27.57±0.2	19.53±0.1	13.48±0.3	56.48±2.8	31.20±0.8

Table 1: Performance Comparison Across Models on the MSMO Dataset

Dataset	Model	R-1 [%]↑	R-2 [%]↑	R-L [%]↑	B-1 [%]↑	B-2 [%]↑	BERTScore [%]↑	M^3AS [%]↑
MSMO	Vision-GPLM	36.92 ±1.7	23.10 ±0.8	31.09 ±0.5	17.38 ±1.2	14.75 ±1.1	34.18 ±2.0	18.58 ±0.9
	VG-GPLMs	40.16±1.6	25.47±1.1	32.09±1.3	20.69±1.1	16.01±0.3	48.94±2.9	23.19±1.2
	Va-SOGM	**41.94±2.1**	**28.15±0.8**	32.60±2.2	**22.32±1.7**	**18.13±0.8**	45.34±2.1	23.43 ±1.6
	ModalSyncSum	39.95±1.7	25.74±1.3	**32.64±1.4**	22.25±1.7	18.03±1.0	**60.71±1.1**	**35.46±1.1**
M3LS	Vision-GPLM	34.91±1.3	21.99±1.1	29.60±1.6	17.64±1.5	14.25±0.6	31.26±2.2	16.10±1.1
	VG-GPLMs	**38.90±1.8**	27.61±0.6	33.74±1.5	**22.81±1.1**	**18.47**±0.3	46.07±2.0	21.25±1.7
	Va-SOGM	38.77±2.4	**28.43±0.9**	**34.35±0.7**	21.75±1.3	18.21±1.4	42.69±1.5	20.11±0.7
	ModalSyncSum	35.19±1.3	26.51±0.5	32.81±1.0	20.81±0.6	17.39±1.1	**59.92±1.9**	**33.69±1.0**
E-Liputan	Vision-GPLM	30.95±0.6	22.42±1.2	27.37±2.1	16.54±1.4	13.15±1.1	35.53±2.1	18.65±1.3
	VG-GPLMs	31.20±1.3	24.28±1.3	29.30±0.3	20.27±1.4	13.50±0.4	48.62±1.4	23.48±1.3
	Va-SOGM	32.92±0.4	27.49±0.6	30.07±1.1	**21.84±1.9**	14.47±1.2	56.16±1.9	23.24±1.4
	ModalSyncSum	**34.66±2.1**	**27.53±1.4**	**32.79±1.7**	21.58±0.6	**15.25±1.2**	**58.68±2.6**	**32.73±0.7**

Table 2: Performance on MSMO, M3LS, and E-Liputan Datasets. ModalSyncSum uses the GLM-4v model by default

Model	R-1↑	R-L↑	B-1↑	BERTScore↑	M^3AS↑
GPT-4o	31.31	25.44	20.15	59.48	–
GPT-4o (Img)	27.22	25.83	17.46	54.37	28.93
w/o ImgC	35.18	29.14	20.49	57.26	30.07
w/o SumC	32.63	28.46	19.95	55.32	33.13
ModalSyncSum	**39.95**	**32.64**	**22.25**	**60.71**	**35.46**

Table 3: Ablation Study Results. Img: image modality; ImgC: image consistency; SumC: summary consistency.

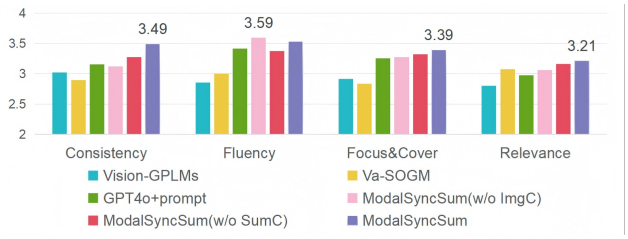

Figure 3: Human Evaluation Experiment

metrics, demonstrating its effectiveness in addressing both linguistic and visual challenges in multimodal summarization.

Human Evaluation

To further evaluate the effectiveness of the proposed framework, we conducted a human evaluation involving 15 graduate students from diverse academic backgrounds, including computer science, journalism, and law. For each model, 500 samples were randomly selected from the MSMO dataset, and participants assessed the generated multimodal summaries—comprising both text and images—using a structured questionnaire. Prior to scoring, all evaluators were briefed on the definitions and criteria for the four evaluation metrics to ensure consistent and objective judgments.

As shown in Figure 3, ModalSyncSum outperforms all baselines across multiple criteria, especially on Consistency—highlighting the effectiveness of its consistency checking. It also excels in Fluency and Focus&Coverage, benefiting from large-scale language models.

Metric Validation: M^3AS

To assess the effectiveness of M^3AS in aligning with human judgments, we evaluate its correlation with four

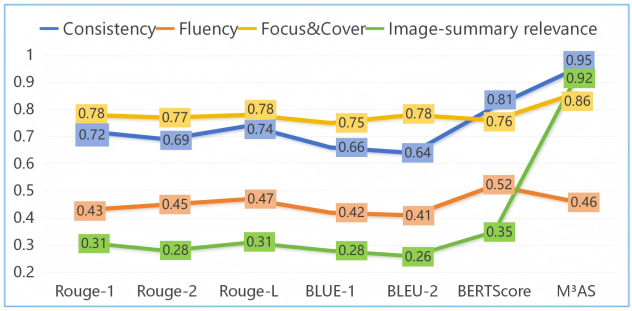

Figure 4: The correlation coefficients between M³AS and four human evaluation metrics: Consistency, Fluency, Focus&Coverage, and Image-summary relevance.

Group	$\text{Score}_{\text{img\_info}}\uparrow$	$\text{Score}_{\text{img\&sum}}\uparrow$	$\text{Score}_{\text{consist}}\uparrow$	$\text{M}^3\text{AS}\uparrow$
G1	13.69	**42.60**	45.36	36.75
G2	11.56	21.33	51.74	34.09
G3	13.60	41.27	24.37	25.90
G4	**15.48**	42.05	**52.51**	**40.63**

Table 4: Ablation study results illustrating the sensitivity of M³AS under four perturbation settings. G1: modifying 1–2 key details in the image description; G2: substituting the original image with an irrelevant one; G3: altering 1–2 key details in the generated summary; G4: using the original, unaltered data.

critical human-centric criteria on 500 manually annotated multimodal summaries. As illustrated in Figure 4, M³AS demonstrates consistently strong alignment with Consistency (0.9588), Image-summary relevance (0.9221), and Focus&Coverage (0.8672), markedly surpassing traditional metrics such as ROUGE, BLEU, and BERTScore. These results indicate that M³AS excels in capturing core aspects of multimodal summary quality, particularly factual accuracy, information completeness, and cross-modal alignment. Although its correlation with Fluency (0.46) is relatively lower, it remains competitive with other metrics, reflecting a greater emphasis on semantic and structural fidelity. Overall, M³AS offers a more comprehensive and human-aligned evaluation of multimodal summaries, underscoring its utility as a robust assessment tool in multimodal settings.

Table **??** shows that each sub-score of M³AS effectively captures different types of perturbations. G4 (original data) achieves the highest scores across all metrics, indicating strong consistency under clean input. G1 leads to a drop in $\text{Score}_{\text{img\_info}}$, showing sensitivity to image description errors. G2 results in a low $\text{Score}_{\text{img\&sum}}$, reflecting disrupted image-summary alignment. G3 significantly lowers $\text{Score}_{\text{consis}}$ and overall M³AS, highlighting the impact of summary inconsistencies. Overall, M³AS demonstrates clear discriminative ability across different perturbation types.

Hyperparameter Search for M³AS

Table 5 presents the performance of ModalSyncSum under different values of τ_1 (the threshold for filtering reference sentences based on image-text similarity) and τ_2 (the threshold for filtering candidate descriptions based on image-description similarity). As τ_1 increases from 0.1 to 0.5 (with τ_2 fixed at the default), ROUGE-1, ROUGE-2, and ROUGE-L scores first remain stable and then decline, while BERTScore and M³AS peak at τ_1 = 0.3, with values of 60.71 and 35.46, respectively. This indicates that increasing τ_1 reduces the number of selected reference sentences, which may lower the lexical overlap between the generated image descriptions and the reference summaries, thus causing a drop in ROUGE scores. When τ_1 is too high, very few sentences are selected as references, significantly reducing the consistency between image descriptions and the source text, which leads to a sharp decline in M³AS.

τ_1	τ_2	R-1↑	R-L↑	BERTScore↑	$\text{M}^3\text{AS}\uparrow$
0.1	*	**36.53**	30.19	57.17	28.80
0.2	*	36.04	**30.48**	55.05	33.08
*	*	34.95	28.64	**60.71**	**35.46**
0.4	*	28.36	23.87	52.91	32.01
0.5	*	27.56	21.09	47.81	25.24
*	0.1	**38.37**	**29.19**	56.30	31.73
*	0.2	36.85	28.54	58.59	31.66
*	0.3	34.08	28.56	60.02	33.45
*	*	34.95	28.64	**60.71**	**35.46**
*	0.5	32.21	27.32	57.33	34.72

Table 5: Ablation results of τ_1 and τ_2 settings. * indicates the default setting (τ_1=0.3, τ_2=0.4).

A similar pattern is observed for τ_2: as it increases from 0.1 to 0.5, ROUGE scores generally decrease, while BERTScore and M³AS peak at τ_2 = 0.4. This suggests that stricter filtering reduces the number of candidate descriptions, potentially lowering lexical overlap and thus ROUGE scores. Excessively high τ_2 values filter out too many candidates, impairing the semantic completeness of image descriptions and negatively impacting M³AS.

Conclusion

To address the hallucination problem and the imbalance between visual and textual attention in large-scale multimodal summarization models, this paper proposes ModalSyncSum, a novel multimodal summarization framework. ModalSyncSum first converts image content into textual descriptions and leverages Named Entity Recognition and Question Answering models to verify the consistency of these descriptions. The framework then fuses the original text with the image descriptions to generate an initial summary, followed by a second-stage consistency verification between the summary and the visual content. Based on this verification, prompt-based refinement is applied to guide the model in correcting factual errors and improving content consistency. Furthermore, we introduce a new evaluation metric, M³AS, which emphasizes semantic relevance and information consistency between text and images, offering a more task-specific and reliable assessment for multimodal summarization. Experimental results across multiple

datasets demonstrate that ModalSyncSum achieves consistently strong performance, confirming its effectiveness and generalizability in multimodal summarization tasks.

Acknowledgements

We would like to express our sincere gratitude to Professor Shengyi Jiang for his invaluable guidance throughout this work. We also extend special thanks to Ziying Rong for her kind assistance. This research was supported in part by the Data Mining Laboratory of the School of Information Science, Guangdong University of Foreign Studies, which provided computational resources for this study.

References

Adrakatti, V. 2024. *Exploring Screen Summarization with Large Language and Multimodal Models*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

ByteDance-Seed Team. 2024. Doubao-Vision-Pro: Multimodal Large Language Model by ByteDance. https://www.doubao.com/chat/. Accessed: 2025-05-05. A proprietary multimodal language model developed by ByteDance.

Colombo, P.; Clavel, C.; and Piantanida, P. 2022. InfoLM: A New Metric to Evaluate Summarization & Data2Text Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 10554–10562.

Fabbri, A. R.; Kryściński, W.; McCann, B.; Xiong, C.; Socher, R.; and Radev, D. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9: 391–409.

Fu, J.; Ng, S.-K.; Jiang, Z.; and Liu, P. 2024. GPTScore: Evaluate as You Desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 6556–6576. Mexico City, Mexico: Association for Computational Linguistics.

Ghosh, A.; Acharya, A.; Jain, R.; et al. 2024. CLIPSyntel: CLIP and LLM Synergy for Multimodal Question Summarization in Healthcare. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 22031–22039.

Google DeepMind. 2024. Gemini: Multimodal Foundation Models from Google. https://gemini.google.com/. Accessed: 2025-05-05.

Jangra, A.; Jatowt, A.; Hasanuzzaman, M.; and Saha, S. 2020. Text-image-video summary generation using joint integer linear programming. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Proceedings, Part II*, 190–198. Lisbon, Portugal: Springer.

Jing, L.; Li, Y.; Xu, J.; Song, Y.; Liu, X.; Yang, Y.; Shen, Y.; and Tong, H. 2023. Vision Enhanced Generative Pre-trained Language Model for Multimodal Sentence Summarization. *Machine Intelligence Research*, 20(2): 289–298.

Koto, F.; Baldwin, T.; and Lau, J. H. 2022. FFCI: A Framework for Interpretable Automatic Evaluation of Summarization. *J. Artif. Int. Res.*, 73.

Krubiński, M.; Ghadery, E.; Moens, M.-F.; et al. 2021. Just Ask! Evaluating Machine Translation by Asking and Answering Questions. In *Proceedings of the Sixth Conference on Machine Translation*, 495–506.

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

Li, Z.; Xu, X.; Shen, T.; Xu, C.; Gu, J.-C.; Lai, Y.; Tao, C.; and Ma, S. 2024. Leveraging Large Language Models for NLG Evaluation: Advances and Challenges. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 16028–16045. Miami, Florida, USA: Association for Computational Linguistics.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 74–81.

OpenAI. 2024. GPT-4o System Card. arXiv:2410.21276.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318.

Pu, X.; Gao, M.; and Wan, X. 2023. Summarization is (Almost) Dead. *arXiv preprint arXiv:2309.09558*.

Song, Y.; Lin, N.; Li, L.; Wu, X.; and Shen, Y. 2024. A Vision Enhanced Framework for Indonesian Multimodal Abstractive Text-Image Summarization. In *Proceedings of the 2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 61–66. IEEE.

Tang, B.; Lin, B.; Chang, Z.; et al. 2024. Multimodal Summarization with Modality Features Alignment and Features Filtering. *Neurocomputing*, 603: 128270.

Verma, Y.; Jangra, A.; Verma, R.; and Saha, S. 2023. Large Scale Multi-Lingual Multi-Modal Summarization Dataset. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Association for Computational Linguistics.

Wang, H.; Liu, J.; Duan, M.; Gong, P.; Wu, Z.; Wang, J.; and Han, B. 2023. Cross-modal knowledge guided model for abstractive summarization. *Neural Computing and Applications*. Published online 27 July 2023.

xAI Team. 2024. Grok-1: A Language Model by xAI. https://x.ai. Accessed: 2025-05-05. A proprietary large language model developed by xAI.

Xiao, M.; Zhu, J.; Zhai, F.; et al. 2025. Pay More Attention to Images: Numerous Images-Oriented Multimodal Summarization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 9379–9392. Mexico City, Mexico: Association for Computational Linguistics.

Yu, T.; Dai, W.; Liu, Z.; and Fung, P. 2023. Vision Guided Generative Pre-trained Language Models for Multimodal Abstractive Summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 3–5. Association for Computational Linguistics.

Zeng, A.; Xu, B.; Wang, B.; and Zhang, C. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. *CoRR*, abs/2406.12793.

Zhang, L.; Zhang, X.; Guo, Z.; et al. 2023. CISum: Learning Cross-Modality Interaction to Enhance Multimodal Semantic Coverage for Multimodal Summarization. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, 370–378. Society for Industrial and Applied Mathematics.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations (ICLR)*.

Zhu, J.; Li, H.; Liu, T.; et al. 2018. MSMO: Multimodal Summarization with Multimodal Output. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4154–4164.