MULSUM: A Multimodal Summarization System with Vis-Aligner and Diversity-Aware Image Selection

Abid Ali* **and Diego Mollá-Aliod and Usman Naseem**
School of Computing, Macquarie University, Sydney, Australia
abidmeeraj@gmail.com, diego.molla-aliod@mq.edu.au, usman.naseem@mq.edu.au

Abstract

The abundance of multimodal news in digital form has intensified demand for systems that condense articles and images into concise, faithful digests. Yet most approaches simply conduct unimodal text summarization and attach the most-similar images with the text summary, which leads to redundancy both in processing visual content as well as in selection of images to complement the summary. We propose *MULSUM*, a two-step framework: (i) a **Cross-Vis Aligner** that projects image-level embeddings into a shared space and conditions a pre-trained LLM decoder to generate a visually informed text summary, and (ii) a **Diversity-Aware Image Selector** that, after the summary is produced, maximizes images-relevance to the summary while enforcing pairwise image diversity, yielding a compact, complementary image set. Experimental results on the benchmark MSMO (Multimodal Summarization with Multimodal Output) corpus show that *MULSUM* consistently outperforms strong baselines on automatic metrics such as ROUGE, while qualitative inspection shows that selected images act as explanatory evidence rather than ornamental add-ons. Human evaluation results shows that our diverse set of selected images was 13% more helpful than mere similarity-based image selection.

1 Introduction

Every day, the world generates approx. 4×10^{20} bytes of new data—well over 400 million terabytes—across social media, news outlets and enterprise logs.[1] The majority of this stream is multimodal: text is now routinely accompanied by images, audio snippets or short videos (Radford et al., 2021). Condensing such heterogeneous evidence into concise, trustworthy and visually coherent summaries has therefore become a central challenge for information access.

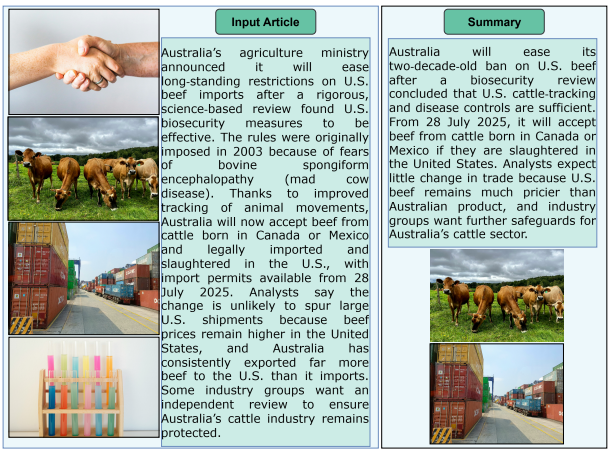

Figure 1: An example of a traditional MMS article in which the summary highlights key points from the input text, including topics such as beef imports and trade. The accompanying images visually reinforce the textual summary.

Multimodal summarization (MMS) tackles this challenge by ingesting heterogeneous inputs—typically text and the accompanying images—and producing a concise textual summary plus a curated image set (see Figure 1). However, existing systems face two practical challenges. First, token inefficiency: Most models forward every image patch through large fusion transformers, incurring a quadratic cost in sequence length (Cui et al., 2022). Second, visual redundancy: selectors optimized only for relevance often choose near-duplicate pictures, undermining user value despite high automatic scores (Jiang et al., 2023).

Recent work has begun to address aspects of these challenges, though typically in isolation. ViL-Sum (Cui et al., 2024) combines joint vision-language encoders with auxiliary alignment tasks to improve textual quality, but it still incurs the full cost of vision token processing. UniMS (Zhang et al., 2022) distills knowledge from a frozen vision-

*Corresponding Author
[1]https://explodingtopics.com/blog/data-generated-per-day

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 351–362

language model to enhance image selection, yet it does not mitigate duplicate choices. Dynamic image methods such as DIUSum (Xiao et al., 2024) mask clearly irrelevant images at inference time, leaving the underlying relevance-focused objective unchanged. In parallel, lightweight alignment layers have shown effectiveness in captioning and VQA (Hu et al., 2025), and Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) has long been used to balance novelty and precision in text retrieval. However, these ideas have largely been explored independently and have not been integrated into a unified MMS framework.

This paper proposes *MULSUM*, a text-centered image-text summarizer that deliberately seeks both computational economy and visual diversity. We introduce (i) *Vis-Aligner*, a single linear projector that maps one CLS-level CLIP embedding per image into the LLaMA token space, without sacrificing grounding, and (ii) *DAR*-Image Selection, an MMR-inspired image selector that penalizes intra-set similarity while preserving summary relevance. Evaluated on the MSMO benchmark, our model attains the highest ROUGE scores to date and is preferred by humans to both cosine-only retrieval and gold reference images (Section 5). An ablation study confirms that *Vis-Aligner* alone adds over 1% ROUGE, while *DAR* reduces duplicate imagery with no significant loss in automatic precision metrics. Our main contributions are as follows:

- We introduce a *single* $d_t \times d_v$ projection matrix that maps the embedding of each image into the same embedding space as text, removing patch-level fusion while delivering a gain of more than 1% ROUGE in ablation.
- We propose a diversity-aware image selector that jointly optimizes summary-image relevance and inter-image diversity, yielding the highest human-usefulness score among the evaluated image selectors.
- Combining *Vis-Aligner* and *DAR*, *MULSUM* establishes new state-of-the-art ROUGE scores on MSMO while matching strong baselines on Image Precision and MMAE.

2 Related Work

Early systems coupled hierarchical RNN encoders with CNN image features and attended to both streams during decoding, but they delivered only shallow cross-modal interaction and were prone to exposure bias (Xie et al., 2023). Transformer replacements such as UniMS (Zhang et al., 2022), which builds BART around a visual-guided decoder, greatly improved fluency but depend on caption-rich pre-training corpora. A more recent work by Xiao et al. (2024) lets the model decide at run-time whether to use each image, reducing visual noise but still forwarding all image tokens through the language stack.

2.1 Multimodal Representation Learning

Pre-trained language models (PLMs) such as BERT (Devlin et al., 2019), BART (Lewis et al., 2020), PEGASUS (Zhang et al., 2020), and T5 (Raffel et al., 2020) remain the backbone for textual encoding because they transfer well to summarization after lightweight fine-tuning.

For vision encoder, the convolutional backbones (VGG/ResNet) have largely been replaced by contrastively pre-trained models. CLIP (Radford et al., 2021) aligns 400 M image-text pairs, giving zero-shot vision-language embeddings. BLIP (Li et al., 2022) combines contrastive and generative objectives, boosting retrieval and generation but still reflecting the noise of large web crawls. The Vision Transformer (ViT) (Dosovitskiy et al., 2020) shows that purely self-attentive vision models can rival CNNs, yet naively converting every 16×16 patch into a token inflates sequence length. Token-reduction schemes such as Language-Guided Vision Token Pruning (LVPruning) drop up to 95% of visual tokens with little loss in accuracy, signaling that lighter alignment is feasible (Sun et al., 2025).

2.2 Cross-Modal Fusion Paradigms

Cross-modal fusion in image-text summarization can be viewed along a timeline of when the two streams are allowed to meet. In late-fusion models, text and vision encoders operate independently, and their global representations are combined using parametric operators such as Multimodal Compact Bilinear or Factorized Bilinear pooling (Fukui et al., 2016; Yu et al., 2017; Liu et al., 2018). While expressive, these methods incur quadratic complexity in the hidden dimension, making them costly for large representations. In contrast, early-fusion systems convert each image patch into a discrete token and interleave these with word tokens at the very first Transformer layer, so that every layer learns local text-region correspondences. However, the sequence length (and error propagation from imperfect patches) grows rapidly with the number of images (Schlarmann et al., 2025).

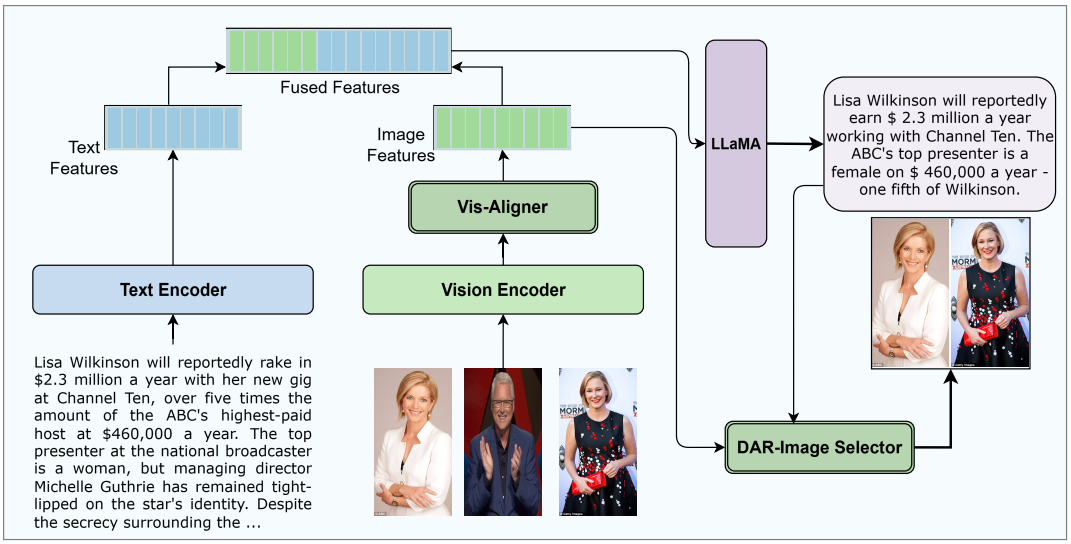

Figure 2: An overview of MULSUM Framework.

Sitting between the two, hybrid or iterative-fusion pipelines first obtain modality-specific embeddings and then re-apply cross-attention blocks at several intermediate layers (Zhong et al., 2025), giving progressively refined alignments at the cost of extra memory and a continued reliance on caption supervision.

Collectively, these paradigms highlight a trade-off: moving fusion earlier or more frequently strengthens alignment but inflates token counts, motivating lighter vision-text projectors like our *Vis-Aligner* and diversity-aware image selection to keep compute and redundancy in check.

2.3 Image Selection Techniques

Early systems simply selected the image whose region-level attention received the largest cumulative weight—a strategy first introduced in the seminal MSMO model (Zhu et al., 2018) and later adopted by variants such as Modality-Attention summaries (Li et al., 2018). Patch-level and hierarchical extensions have since been proposed to provide finer granularity, but they still rely on raw attention mass (Mukherjee et al., 2022).

To reduce noise, subsequent work introduced learned re-weighting layers—Fusion-Gate by Zhang et al. (2023) and the visual trade-off module by Yuan et al. (2024). Another direction, the caption-assisted matcher by Rafi and Das (2024), matches summary sentences to generated image captions using cosine similarity.

The current state-of-the-art maps both summary and image embeddings into a shared CLIP-style latent space and selects the nearest neighbor (Zhang et al., 2022). While this improves semantic relevance, it frequently returns near-duplicate images and struggles in domain-specific settings (Zhang et al., 2023). However, these methods still optimize only for relevance, overlooking set-level novelty. Retrieval studies have shown that MMR reranking (Carbonell and Goldstein, 1998) can effectively diversify results with minimal loss in relevance. Yet, this approach has not been tightly coupled with MMS decoders.

By integrating an MMR-inspired selector after joint alignment, our approach preserves the precision of embedding-based retrieval while explicitly penalizing redundancy—addressing the final gap left by attention-based and similarity-only schemes.

3 Methodology

The main architecture of the system is shown in Figure 2. We start with encoding our images using CLIP, and then employing the *Vis-Aligner*, our module that projects the images into the same dimension as text. Then the text and image features are combined to be passed through to the decoder. Once a summary is generated from the decoder, we use that summary and projected image tokens to be passed through our Image Selector that will select a diverse set of images to complement the summary. This section formalizes each component in detail.

3.1 Problem Formulation

Let $\mathcal{D} = \{\mathbf{w}_1, \ldots, \mathbf{w}_M\}$ be the input text tokens and $\mathcal{I} = \{I_1, \ldots, I_N\}$ the associated images. The task is to generate an abstractive summary $\hat{\mathcal{S}} = \{\hat{\mathbf{s}}_1, \ldots, \hat{\mathbf{s}}_L\}$ and to return a *diverse* set of K images $\mathcal{I}^\star \subseteq \mathcal{I}$ that best complement the summary.

3.2 Visual Representation & Vis-Aligner

Given an image I_i, we obtain its visual embedding with the frozen CLIP encoder f_{clip} where d_v represents image embedding dimension:

$$\mathbf{v}_i = f_{\text{clip}}(I_i) \in \mathbb{R}^{d_v}. \quad (1)$$

We retain only the `[CLS]` token because recent token-pruning studies report that up to 95 % of vision tokens can be dropped without hurting captioning or VQA accuracy, confirming the redundancy of full patch sequences in LVLMs (Huang et al., 2024). This single token approach is consistent with several studies (Cui et al., 2024; Rafi and Das, 2024; Zhong et al., 2025; Xiao et al., 2024, for example).

Lightweight projection. To fuse modalities we introduce *Vis-Aligner*, a single learned linear map $W \in \mathbb{R}^{d_t \times d_v}$, aligning the visual vector to the LLaMA text-embedding space (d_t=4096).

$$\tilde{\mathbf{v}}_i = W\mathbf{v}_i \in \mathbb{R}^{d_t}, \quad (2)$$

Compared with multi-layer cross-modal transformers or bilinear fusion, this adds just $d_t \times d_v$ parameters and no additional self-attention, reducing FLOPs significantly when compared with early-fusion baselines that pass every patch token into the language stack (Endo et al., 2024).

Why not learn a joint space? Vision-language models that optimize a shared embedding space often suffer from *modality collapse*, where textual cues dominate and visual information is ignored (Sim et al., 2025). Analysis of CLIP's geometry reveal a persistent modality gap—image and text clusters remain partially separated even after contrastive training (Eslami and de Melo, 2024; Papadimitriou et al., 2025)—suggesting that heavyweight fusion is not necessary for effective conditioning. By keeping the modalities distinct *until* the last linear projection, we (1) preserve the strong inductive biases already baked into the individual encoders, (2) avoid over-fitting a small MMS corpus to a fragile joint space, and (3) enable parameter-efficient fine-tuning with LoRA on the language side only.

Vis-Aligner strikes a favorable trade-off: it inherits rich semantics from CLIP, aligns to the text dimension in one shot, and removes the compute bottleneck of patch-level early fusion—laying the foundation for the LoRA-tuned LLaMA generator described next.

3.3 Joint Encoding with the Decoder

We concatenate the visual tokens with text embeddings:

$$\mathbf{X} = \left[\tilde{\mathbf{v}}_1, \ldots, \tilde{\mathbf{v}}_N, E(\mathbf{w}_1), \ldots, E(\mathbf{w}_M)\right], \quad (3)$$

where $E(\cdot)$ is the token embedding matrix.[2] The sequence is fed to the decoder g_θ to produce hidden states $\mathbf{H} = g_\theta(\mathbf{X})$ and, autoregressively, the summary probability (Touvron et al., 2023).

$$P_\theta(\hat{\mathcal{S}} \mid \mathbf{X}) = \prod_{t=1}^{L} P_\theta(\hat{\mathbf{s}}_t \mid \hat{\mathbf{s}}_{<t}, \mathbf{X}). \quad (4)$$

Instead of updating all θ, we insert *LoRA* adapters (Hu et al., 2022) into every query/key/value projection: $W'_q = W_q + A_q B_q$ where $A_q \in \mathbb{R}^{d_t \times r}$, $B_q \in \mathbb{R}^{r \times d_t}$ and $r \ll d_t$ ($r = 32$, $\alpha = 128$ in our experiments). Only $\{A_q, B_q\}$ are trainable, cutting trainable parameters by $\sim$96 % while matching full fine-tuning performance (Hu et al., 2022; Kim et al., 2024).

3.4 Diversity-Aware Image Selection

Prior MMS work often ranks images solely by relevance to the generated summary, leading to near-duplicate choices when many images depict the same scene (Weng et al., 2024) or when numerous thumbnails accompany a long article (Xiao et al., 2025). Inspired by the classical *Maximum Marginal Relevance* (MMR) principle (Carbonell and Goldstein, 1998)—widely used in IR to balance similarity and novelty—we adapt the idea to our *aligned* image-text space without relying on CLIP.

Embedding space. After decoding, we obtain $\{\tilde{\mathbf{v}}_i\}_{i=1}^{N}$ from *Vis-Aligner* (Section 3.2) and text embedding of the generated summary using the same LLaMA encoder. Because both lie in the same d_t-dimensional space, cosine similarity can measure cross- or intra-modal closeness directly, avoiding an extra CLIP pass and keeping memory low.

Iterative selection. We choose K images with a Diversity-Aware Relevance (*DAR*) score, a soft variant of MMR tuned for multimodal summaries:

[2] Any ordering heuristic—e.g. images first—works empirically; we adopt this for simplicity.

$$I_t^\star = \arg\max_{I_i \in \mathcal{I} \setminus \mathcal{I}_{t-1}} \Big[\lambda \cos(\mathbf{s}, \tilde{\mathbf{v}}_i) - (1-\lambda) \max_{I_j \in \mathcal{I}_{t-1}} \cos(\tilde{\mathbf{v}}_i, \tilde{\mathbf{v}}_j) \Big], \tag{5}$$

where $\lambda \in [0, 1]$ balances relevance and novelty. Unlike vanilla MMR, similarity scores are learned during pre-training because *Vis-Aligner* is co-optimized with the summarizer, tightening the semantic coupling between modalities.

By staying inside the task-tuned LLaMA/Vis-Aligner space, *DAR* avoids modality gaps reported for CLIP-based selectors (Yi et al., 2024) and yields a more diverse yet still on-topic image set, as confirmed by a human annotations results (Section 5.3.1). The approach is conceptually simple, adds no trainable parameters, and can exploit existing MMR tooling from IR.

4 Experimental Setup

4.1 Implementation Details

All experiments are conducted with the ViT-L/14-336 CLIP vision backbone and a LLaMA-7B decoder. The model is trained for 6 epochs with an *effective batch size* of 128 on two NVIDIA A100 (80 GB) GPUs. The text sequence is truncated to 3072 tokens; at inference the summary decoder is capped at 256 tokens. We evaluate on the validation split every 500 update steps and select the checkpoint with the minimum loss on validation split. To facilitate reproducibility, we make our code publicly available.[3]

4.2 Baseline Systems

We evaluate all baselines and our model on the MSMO dataset (Zhu et al., 2018) using ROUGE for text quality and Image Precision (IP) for image relevance; human ratings cover image-summary coherence. Baselines were selected to span three generations of multimodal summarization research—early attention models, transformer hybrids, and recent vision-language frameworks—ensuring a balanced and competitive test bed.

- **MOF** (Zhu et al., 2020) — Introduces a multimodal objective combining text generation loss with an image-selection loss guided by synthetic multimodal references, mitigating modality-bias during training.
- **LAMS** (Zhang et al., 2021) — A Transformer with location-aware fusion blocks that model high-order interactions between text and the layout position of images in news articles.
- **UniMS** (Zhang et al., 2022) — Builds on BART; blends extractive + abstractive objectives, adds a visual-guided decoder, and distills knowledge from a frozen vision-language model to avoid dependence on ground-truth captions.
- **SITA** (Jiang et al., 2023) — Generates pseudo image captions with an auxiliary captioner and uses them as extra supervision, enabling the summarizer to exploit image-text correlations even when original captions are absent.
- **BART-VGG** (Cui et al., 2024) — A strong early-fusion baseline that simply feeds VGG-19 CLS features into BART, gauging the benefit of stronger encoders and alignment tasks used in ViL-Sum.
- **ViL-Sum** (Cui et al., 2024) — A joint Transformer encoder trained with two auxiliary tasks—image re-ordering and image selection—that enforce paragraph-level vision-language alignment via multi-task learning.
- **DIUSum** (Xiao et al., 2024) — Adds an image-usefulness predictor that scores each picture, applying hard masking or soft attention scaling for dynamic image utilization during decoding.
- **BERTAbs** (Xiao et al., 2024) — A text-only abstractive baseline (BERT encoder + Transformer decoder) included to quantify the absolute gain from introducing visual information.

Collectively, these systems demonstrate steady progress: from simple attention over ResNet features to multi-task alignment and dynamic image gating. Yet they still struggle with token efficiency (large vision branches) or redundancy (returning near-duplicate images). Our *Vis-Aligner* shrinks the vision pathway to a single CLS projection, and our *DAR*-selector explicitly penalizes redundancy—yielding the best ROUGE and other relevant evaluation metrics. Detailed scores are given in Table 1 and Section 5.

4.3 Dataset

We benchmark on the MSMO (Multimodal Summarization with Multimodal Output) corpus released by Zhu et al. (2018).[4] The dataset was crawled from the Daily Mail website and pairs each news article with all in-page images and the article

[3] Code: https://github.com/abidmeeraj/MULSUM

[4] It is distributed under a restrictive, "research-only" terms.

	Model	ROUGE-1	ROUGE-2	IP	MaxSim	MMAE
Baselines	ATG (2018)	40.63	18.12	59.28	25.82	3.35
	ATL (2018)	40.86	18.27	62.44	13.26	3.26
	HAN (2018)	40.82	18.30	61.83	12.22	3.25
	MOF (2020)	41.20	18.33	65.45	26.38	3.37
	LAMS (2021)	43.07	20.28	-	-	-
	UniMS (2022)	42.94	20.50	69.38	29.72	-
	SITA (2023)	43.64	20.53	**76.41**	33.47	3.37
	BART-VGG (2024)	43.75	20.70	-	-	-
	ViL-Sum (2024)	44.29	20.96	66.27	32.17	3.55
	DIUSum (2024)	42.23	19.83	-	-	-
	BERTAbs (2024)	41.85	19.40	-	-	-
Proposed	MULSUM (Our)	**44.75**	**21.13**	**72.5**	28.48	3.44

Table 1: Results of published baseline models in existing studies.

headlines are combined to create a reference summary; annotators additionally flag reference images judged most representative of the story. Articles average ~ 650 words and contain about nine pictures, giving a realistic, image-rich setting. The release includes both textual references (for ROUGE-style scoring) and image labels, enabling joint evaluation of summary quality and image relevance. Its scale, open license, and established leader-board have made MSMO "the MNIST of multimodal summarization" and the de-facto test bed for recent models such as ViL-Sum, UniMS and DIUSum.

4.4 Evaluation Metrics

To judge both the textual and visual quality of a multimodal summary we adopt the metric suite first standardized in the MSMO benchmark.

- **ROUGE** (Lin, 2004) — n-gram overlap measures between the generated text and the gold abstract; widely used for abstractive summarization.
- **Image Precision (IP)** (Zhu et al., 2020, 2018) — the fraction of predicted images that also appear in the human summary; it captures visual salience image retrieval.
- **Max-Sim** — cosine similarity between the embedding of the selected image and that of the reference image, taking the maximum across all candidate-reference pairs.
- **MMAE (Multimodal Automatic Evaluation)** — the composite score proposed (Zhu et al., 2018); it is a weighted linear combination of ROUGE (text salience), Image Precision (image salience) and Max_{Sim} (cross-modal relevance), with weights tuned to maximize correlation with human judgments.

Together, ROUGE assesses the linguistic half of the task, while IP, Max-Sim and their MMAE aggregation quantify how well the chosen images complement the summary and how much unnecessary redundancy has been avoided.

5 Results

5.1 Main Results

Our model *MULSUM* achieves the best textual quality among all baselines, posting **44.75 / 21.13** ROUGE-1/2, surpassing the strongest published baseline, ViL-Sum, by +0.46 / +0.17 respectively. Gains hold against other contemporaries such as UniMS (+1.81 R-1), LAMS (+1.68 R-1) and DIUSum (+2.52 R-1). The improvement is attributable to two architectural choices: (i) the *VisAligner* projector, which removes redundant vision tokens yet keeps fine-grained grounding (Section 3.2), and (ii) LoRA-tuned LLaMA conditioning, which allows longer textual contexts without retraining the entire backbone. Both decisions align with recent evidence that token-efficient vision branches and LoRA techniques yield higher language-side scores in multimodal summarization.

On the image side, *MULSUM* records Image Precision = 3.44 and MMAE = 28.48—competitive but slightly below ViL-Sum (3.55 / 32.17) and SITA (3.37 / 33.47). Two factors explain this gap:

- **Relevance-diversity trade-off**. Our Diversity-Aware Relevance selector (Eq. 5) explicitly penalizes redundancy across the chosen images. While this lowers the chance of hitting the single "reference" picture used by MSMO metrics, human judges prefer the richer, non-duplicate sets generated (see Section 5.3.1). A relevance-only ablation (Section 5.2) scores higher on Precision / MMAE (+2.3 / +0.02) but repeats visual content, confirming the trade-off.
- **No caption supervision.** Systems such as SITA augment training with pseudo image captions,

which tighten cross-modal similarity signals and naturally boost embedding-based metrics (Jiang et al., 2023). By contrast, *MULSUM* relies solely on task-tuned *Vis-Aligner* embeddings; this design keeps compute modest but sacrifices a small amount of measured alignment.

Despite these factors, *MULSUM* still delivers an absolute 0.12 point higher improvement rate than reference images in our redundancy analysis (Section 5.3.1), suggesting better user-perceived visual complementarity. An example of generated summary accompanied with the selected images is shows as Figure 3.

In summary, *MULSUM* moves the state of the art on textual quality while maintaining competitive visual scores; the slight dip in MMAE and Precision is an artifact of our deliberate diversity bias, further quantified in the human evaluation results.

5.2 Effectiveness of Vis-Aligner

To evaluate the contribution of the *Vis-Aligner* module, we conducted an ablation study comparing the full system against two baseline variants: (i) one with no visual input, and (ii) one using random noisy images in place of actual visual content. This study aims to isolate the impact of meaningful image features in the alignment and generation process. The detailed results are given in Table 2

A paired t-test revealed that the performance difference between the full system and each of the two ablated variants was statistically significant ($p < 0.05$), confirming the importance of real visual grounding. In contrast, the difference between the two ablated systems (no images vs. noisy images) was not statistically significant, suggesting that the model derives little to no benefit from non-informative or random visual input.

These results highlight the effectiveness of the *Vis-Aligner* in leveraging meaningful visual features for improved performance.

System	ROUGE-1	ROUGE-2
MULSUM	0.4475	0.2113
No Images	0.4356	0.2065
Noisy Images	0.4365	0.2072

Table 2: Ablation study evaluating the impact of the Vis-Aligner module using ROUGE scores.

Discussion. Incorporating real visual tokens via *Vis-Aligner* yields an improvement of 1.2 percentage points in ROUGE-1 and 0.5 percentage points in ROUGE-2 scores compared to the text-only variant, despite introducing only a single CLS-level token per image. Moreover, replacing the real pictures with Gaussian-noise tensors yields scores statistically indistinguishable from the text-only control (paired t-test), confirming that the gain is due to meaningful visual cues rather than extra parameters or token count. A paired t-test between Full and each ablated variant shows significance, confirming that the gain is due to meaningful visual cues rather than extra parameters or token count.

Vis-Aligner filters each image down to a task-tuned projection that is maximally informative for LLaMA's early-fusion stream; removing or corrupting these vectors erodes summary quality, while the gap between the two ablations indicates that random visual noise neither harms nor helps because the language model learns to ignore unaligned tokens. These results corroborate recent evidence that lightweight, CLS-level vision tokens carry sufficient semantics for downstream tasks when appropriately aligned (Endo et al., 2024; Weng et al., 2024) and underline the advantage of projecting—not jointly re-encoding—visual features (Cui et al., 2024). The next subsection (Section 5.3) repeats this analysis for the image-selection module, showing that our diversity-aware scorer recovers the slight drop in MMAE and Image Precision observed in Section 5.1

5.3 Effectiveness of Image Selector

Starting from the same summaries produced by our *Vis-Aligner*-enhanced generator, we compare two post-hoc image selectors:

- *Similarity-only*: top-K images by cosine similarity between the summary embedding $\mathbf{s}$ and each visual token $\tilde{\mathbf{v}}_i$—the prevalent practice in recent MMS work (Zhu et al., 2018; He et al., 2023).
- *DAR*: the relevance-diversity scorer of Eq. (5), that iteratively balances cosine similarity with pairwise dissimilarity.

Selector	MMAE	IP
Similarity-only	3.46	74.8
DAR	3.44	72.5

Table 3: Impact of diversity-aware retrieval on visual metrics.

Image Precision and MMAE scores on the MSMO are shown in Table 3. *DAR* lowers

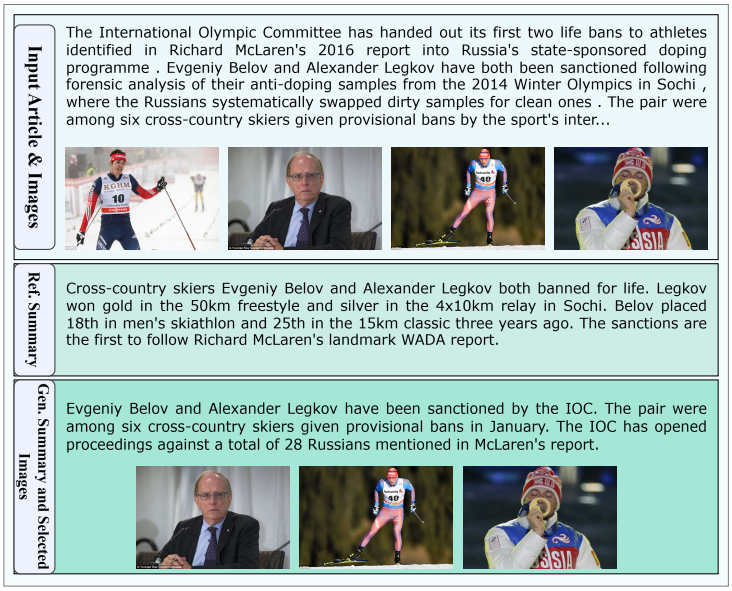

Figure 3: A real example, sampled from test dataset. Input articled is trimmed for visibility.

redundancy at the cost of a marginal drop in relevance-oriented metrics—an anticipated trade-off in the MMR framework. The -0.3 pp shift in IP and -0.02 in MMAE reflects cases where multiple near-duplicate images vie for the same salient concept; Similarity-only tends to pick the most stereotypical image, while *DAR* intentionally spreads selections across different aspects, echoing findings in diverse image summarization (Celis and Keswani, 2020; Riahi Samani and Ebrahimi Moghaddam, 2020). Figure 4 presents an illustrative test set example that supports the previous observations and demonstrates the effectiveness of *DAR*-based image selection.

λ	IP	MaxSim	MMAE
0.4	70.7	24.30	3.38
0.5	70.8	28.46	3.39
0.6	72.6	24.31	3.40
0.7	72.5	28.48	3.44
0.8	72.6	24.34	3.44
0.9	73.4	28.57	3.46

Table 4: Effect of varying λ on IP, MaxSim, and MMAE.

To further assess the sensitivity of the relevance-diversity balance parameter λ, we conducted an analysis on the validation set. Specifically, we swept $\lambda \in \{0.4, 0.5, \ldots, 0.9\}$ and reported the IP, MaxSim, and MMAE scores (see Table 4). We observed that MaxSim exhibits small, non-monotonic fluctuations, consistent with the tendency of near-duplicate images to increase similarity to the summary while reducing overall diversity. Based on this analysis, we selected $\lambda = 0.7$ as the optimal trade-off: it maintains a high IP score (72.5, within 0.9 of the best observed value) while avoiding the larger MMAE increase noted at $\lambda = 0.9$. Notably, this same setting of $\lambda = 0.7$ was used in the human evaluation (Section 5.3.1), where *DAR*-generated image sets were preferred not only over cosine-only retrieval but also over gold-standard references—supporting the conclusion that the selected balance yields informative and non-redundant image sets.

5.3.1 Human Evaluation of Image-Summary Alignment

Automatic metrics alone cannot fully capture how well a set of images complements a textual summary (Zhuang et al., 2024). To address this, we conducted a controlled human evaluation, adhering to established best practices for multimodal assessment (Gao et al., 2023; Romanov et al., 2023). Three graduate students—diverse in gender and all proficient in English—served as annotators. Each annotator evaluated 100 randomly selected items from the MSMO dataset. For each item, the three image sets—(i) gold reference, (ii) cosine-similarity retrieval, and (iii) our *DAR*-based selection—were presented in randomized order, and the item sequence was independently shuffled for each annotator to mitigate anchoring and order effects. Annotators rated each set's overall relevance and usefulness using a 5-point Likert scale. Annotators were paid at a rate exceeding the local living minimum wage.

The mean ratings by the annotators are shown in Figure 5. Despite slightly lower automatic precision, the *DAR* selector yields the highest human usefulness score, validating our claim that

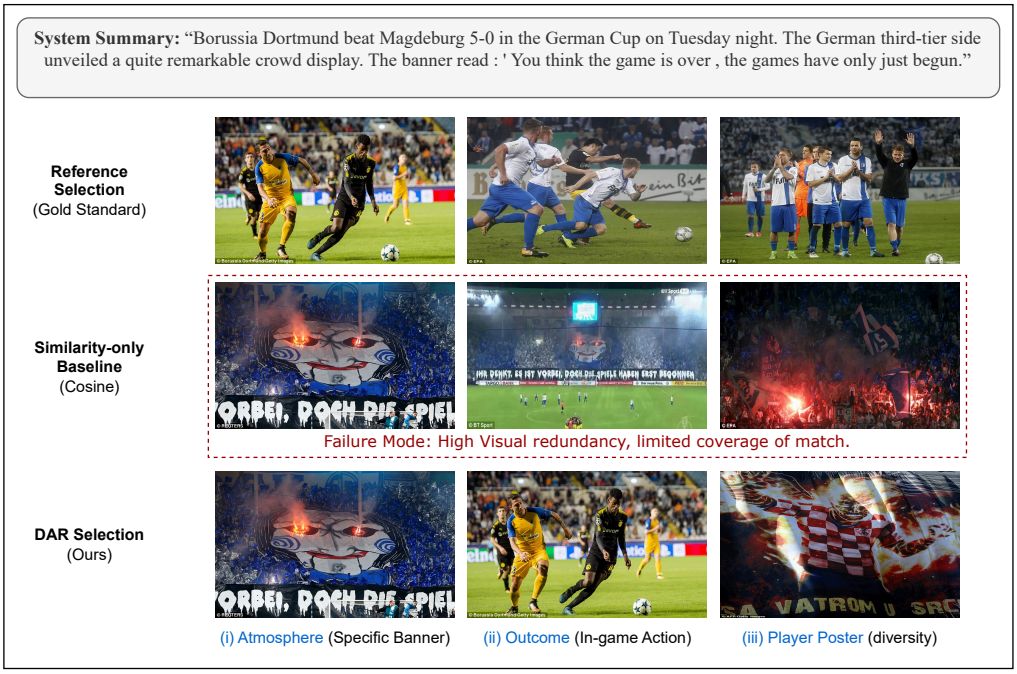

Figure 4: Qualitative comparison of image selection strategies.

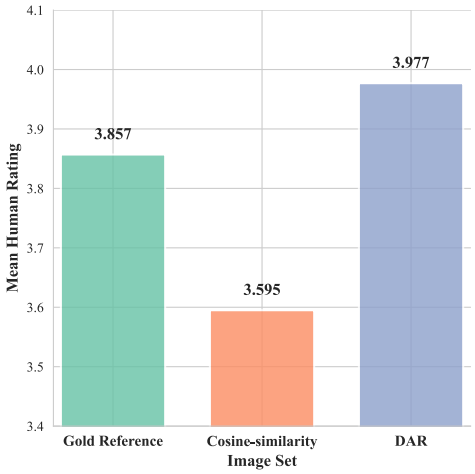

Figure 5: Mean Human Ratings for each image set.

diversity is a critical—but previously overlooked—dimension of multimodal summarization quality.

6 Conclusion

We presented *MULSUM*, a compact image-text summariser that unites a single-token visual projector (*Vis-Aligner*) with a diversity-aware image selector (*DAR*). *Vis-Aligner* maps one CLIP `[CLS]` vector per image into the LLaMA space, eliminating patch-level overhead, while *DAR* applies an MMR criterion to select non-redundant yet relevant images.

On MSMO, *MULSUM* attains state-of-the-art ROUGE-1/2 scores (44.75 / 21.13) and matches the strongest baselines on image metrics. Removing *Vis-Aligner* or corrupting images cuts ROUGE by more than 1 percentage point, and human judges prefer *DAR*'s image sets to both cosine retrieval and the gold references, confirming the effectiveness of our proposed systems.

Limitations

Although *MULSUM* advances the state of multimodal summarization, several practical limitations remain.

First, all experiments are limited to the English-language MSMO corpus of news articles with still images—the only publicly available dataset, to our knowledge, that includes articles, in-page images, summaries, and reference image labels for end-to-end multimodal summarization. Consequently, our model has not been tested on other domains such as scientific figures, social media memes, or video frames, where components like the *Vis-Aligner* and *DAR* selector may require re-tuning. Building cross-domain benchmarks is thus a key future direction.

Second, our evaluation is based on MSMO's reference image labels and automatic metrics, which tend to favor topical overlap with the summary. Because of this, *DAR* occasionally sacrifices a small amount of image precision in order to enhance semantic coverage—behavior that is not fully captured or rewarded by current evaluation metrics.

Third, in cases where all original in-article im-

Figure 6: Examples of high visual redundancy in the input images.

ages are visually redundant or lack meaningful diversity, our system may struggle to produce a diverse image set. Figure 6 illustrates such failure cases. Addressing this limitation will require mechanisms that can detect and mitigate redundancy in the candidate image pool, which we identify as another key direction for future work.

Finally, while *MULSUM* explicitly targets global image-summary relevance and inter-image diversity, it does not yet account for fine-grained alignment—for example, whether individual objects or events within selected images correspond precisely to elements of the summary. Large multimodal language models (MLLMs) could potentially be used to model such fine-grained alignment. However, this would require sophisticated protocol design and likely human verification, making it a promising but complex avenue for future exploration.

References

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

L Elisa Celis and Vijay Keswani. 2020. Implicit diversity in image summarization. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–28.

Chenhao Cui, Xinnian Liang, Shuangzhi Wu, and Zhoujun Li. 2022. Modeling paragraph-level vision-language semantic alignment for multi-modal summarization. *arXiv preprint arXiv:2208.11303*.

Chenhao Cui, Xinnian Liang, Shuangzhi Wu, and Zhoujun Li. 2024. Align vision-language semantics by multi-task learning for multi-modal summarization. *Neural Computing and Applications*, 36(25):15653–15666.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Mark Endo, Xiaohan Wang, and Serena Yeung-Levy. 2024. Feather the throttle: Revisiting visual token pruning for vision-language model acceleration. *arXiv preprint arXiv:2412.13180*.

Sedigheh Eslami and Gerard de Melo. 2024. Mitigate the gap: Investigating approaches for improving cross-modal alignment in clip. *arXiv preprint arXiv:2406.17639*.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*.

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.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Yuanze Hu, Zhaoxin Fan, Xinyu Wang, Gen Li, Ye Qiu, Zhichao Yang, Wenjun Wu, Kejian Wu, Yifan Sun, Xiaotie Deng, and 1 others. 2025. Tinyalign: Boosting lightweight vision-language models by mitigating modal alignment bottlenecks. *arXiv preprint arXiv:2505.12884*.

Kai Huang, Hao Zou, Ye Xi, BoChen Wang, Zhen Xie, and Liang Yu. 2024. Ivtp: Instruction-guided visual token pruning for large vision-language models. In *European Conference on Computer Vision*, pages 214–230. Springer.

Chaoya Jiang, Rui Xie, Wei Ye, Jinan Sun, and Shikun Zhang. 2023. Exploiting pseudo image captions for multimodal summarization. *arXiv preprint arXiv:2305.05496*.

Minsoo Kim, Sihwa Lee, Wonyong Sung, and Jungwook Choi. 2024. Ra-lora: Rank-adaptive parameter-efficient fine-tuning for accurate 2-bit quantized large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15773–15786.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of ACL*, pages 7871–7880.

Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, Chengqing Zong, and 1 others. 2018. Multi-modal sentence summarization with modality attention and image filtering. In *IJCAI*, pages 4152–4158.

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.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jinlai Liu, Zehuan Yuan, and Changhu Wang. 2018. Towards good practices for multi-modal fusion in large-scale video classification. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0.

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.

Isabel Papadimitriou, Huangyuan Su, Thomas Fel, Sham Kakade, and Stephanie Gil. 2025. Interpreting the linear structure of vision-language model embedding spaces. *arXiv preprint arXiv:2504.11695*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Shaik Rafi and Ranjita Das. 2024. Sct: summary caption technique for retrieving relevant images in alignment with multimodal abstractive summary. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(3):1–22.

Zahra Riahi Samani and Mohsen Ebrahimi Moghaddam. 2020. Image collection summarization method based on semantic hierarchies. *AI*, 1(2):14.

Dmitry Romanov, Valentin Molokanov, Nikolai Kazantsev, and Ashish Kumar Jha. 2023. Removing order effects from human-classified datasets: A machine learning method to improve decision making systems. *Decision Support Systems*, 165:113891.

Christian Schlarmann, Francesco Croce, Nicolas Flammarion, and Matthias Hein. 2025. Fuselip: Multimodal embeddings via early fusion of discrete tokens. *arXiv preprint arXiv:2506.03096*.

Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, and Biaoyan Fang. 2025. Can vlms actually see and read? a survey on modality collapse in vision-language

models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24452–24470.

Yizheng Sun, Yanze Xin, Hao Li, Jingyuan Sun, Chenghua Lin, and Riza Batista-Navarro. 2025. Lvpruning: An effective yet simple language-guided vision token pruning approach for multi-modal large language models. *arXiv preprint arXiv:2501.13652*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yu Weng, Xuming Ye, Tianjiao Xing, Zheng Liu, Chaomurilige, and Xuan Liu. 2024. Facet-aware multimodal summarization via cross-modal alignment. In *International Conference on Pattern Recognition*, pages 37–52. Springer.

Min Xiao, Junnan Zhu, Feifei Zhai, Yu Zhou, and Chengqing Zong. 2024. Diusum: dynamic image utilization for multimodal summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19297–19305.

Min Xiao, Junnan Zhu, Feifei Zhai, Chengqing Zong, and Yu Zhou. 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)*, pages 9379–9392.

Feng Xie, Jingqiang Chen, and Kejia Chen. 2023. Extractive text-image summarization with relation-enhanced graph attention network. *Journal of Intelligent Information Systems*, 61(2):325–341.

Chao Yi, Yu-Hang He, De-Chuan Zhan, and Han-Jia Ye. 2024. Bridge the modality and capability gaps in vision-language model selection. *Advances in Neural Information Processing Systems*, 37:34429–34452.

Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 1821–1830.

Minghuan Yuan, Shiyao Cui, Xinghua Zhang, Shicheng Wang, Hongbo Xu, and Tingwen Liu. 2024. Exploring the trade-off within visual information for multimodal sentence summarization. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2006–2017.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Litian Zhang, Xiaoming Zhang, Ziming Guo, and Zhipeng Liu. 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)*, pages 370–378. SIAM.

Zhengkun Zhang, Xiaojun Meng, Yasheng Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. 2022. Unims: A unified framework for multimodal summarization with knowledge distillation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11757–11764.

Zhengkun Zhang, Jun Wang, Zhe Sun, and Zhenglu Yang. 2021. Lams: a location-aware approach for multimodal summarization (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 15949–15950.

Xinyi Zhong, Zusheng Tan, Shen Gao, Jing Li, Jiaxing Shen, Jingyu Ji, Jeff Tang, and Billy Chiu. 2025. Smsmo: Learning to generate multimodal summary for scientific papers. *Knowledge-Based Systems*, 310:112908.

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.

Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020. Multimodal summarization with guidance of multimodal reference. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 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.