Enhancing Large Language Models for Scientific Multimodal Summarization with Multimodal Output

Zusheng Tan[1]*, Xinyi Zhong[1]*, Jing-Yu Ji[1], Wei Jiang[2], Billy Chiu[1]†

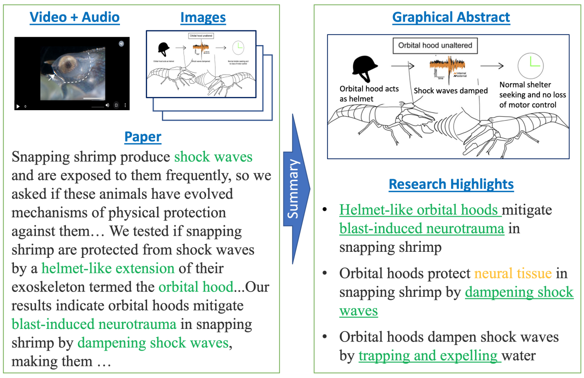
[1]School of Data Science, Lingnan University
[2]School of Economics, Qingdao University
allentan@ln.hk, {xinyizhong, jingyuji}@ln.edu.hk, xy072281@pku.edu.cn, billychiu@ln.edu.hk

Abstract

The increasing integration of multimedia such as videos and graphical abstracts in scientific publications necessitates advanced summarization techniques. This paper introduces Uni-SciSum, a framework for *Scientific Multimodal Summarization with Multimodal Output* (SMSMO), addressing the challenges of fusing heterogeneous data sources (e.g., text, images, video, audio) and outputting multimodal summary within a unified architecture. Uni-SciSum leverages the power of large language models (LLMs) and extends its capability to cross-modal understanding through *BridgeNet*, a query-based transformer that fuses diverse modalities into a fixed-length embedding. A two-stage training process, involving modal-to-modal pre-training and cross-modal instruction tuning, aligns different modalities with summaries and optimizes for multimodal summary generation. Experiments on two new SMSMO datasets show Uni-SciSum outperforms uni- and multi-modality methods, advancing LLM applications in the increasingly multimodal realm of scientific communication.

1 Introduction

Scientific publications are getting more “multimedia”, containing not only text but also visual and auditory content. A popular multimedia publication format nowadays comprises a presentation video, as well as the corresponding Graphical Abstracts (GA), which serve as a diagrammatic summary, and text-based Research Highlights (see Figure 1). The GA helps readers gain a visualized understanding of the paper, while the text offers more detailed explanations. By combining information from different modalities, summaries become more accurate and effectively convey the paper’s main message. This highlights the need for SMSMO (*Scientific Multimodal Summarization with Multimodal Output*) systems capable of generating multimodal summaries from various sources, streamlining the reading process for both editors and readers.


Video + Audio
Images
Orbital hood unaltered
Orbital hood acts as helmet
Shock waves damped
Normal shelter seeking and no loss of motor control
Paper
Summary
Graphical Abstract
Orbital hood unaltered
Orbital hood acts as helmet
Shock waves damped
Normal shelter seeking and no loss of motor control
Research Highlights

Snapping shrimp produce shock waves and are exposed to them frequently, so we asked if these animals have evolved mechanisms of physical protection against them… We tested if snapping shrimp are protected from shock waves by a helmet-like extension of their exoskeleton termed the orbital hood...Our results indicate orbital hoods mitigate blast-induced neurotrauma in snapping shrimp by dampening shock waves, making them …

- Helmet-like orbital hoods mitigate blast-induced neurotrauma in snapping shrimp
- Orbital hoods protect neural tissue in snapping shrimp by dampening shock waves
- Orbital hoods dampen shock waves by trapping and expelling water

Figure 1: A paper-summary example taken from our $\text{SMSMO}_{\text{Cellpress}}$ dataset. The green words in the text summary represent keywords that exist in the source text, whereas the yellow words represent concepts described in video/audio/images. Underlined words represent items that presented across multiple modalities.

In SMSMO, the challenges are two-fold. On the one hand, the heterogeneity of SMSMO data sources, encompassing text, images, video, and audio, presents a challenge in effectively fusing these diverse elements. On the other hand, current scientific summarization frameworks are mainly optimized on modality-specific blocks (Atri et al., 2021, 2023; Kumar et al., 2024), which restricts their applicability to specific data modalities. Models once trained on, for example, *text+video* pairs, there is no straightforward way to apply them to *text+image* or *text-only* data.

Large Language Models (LLMs) have demonstrated remarkable capabilities in various text-based scientific Natural Language Processing (NLP) tasks (Beltagy et al., 2019a, 2020; Guo et al., 2022; Zhang et al., 2020), offering a potential foundation for multimodal summarization. However, effectively integrating multimodal information into these LLMs for SMSMO remains an open chal-

*Equal contribution
†Corresponding author

Proceedings of the 31st International Conference on Computational Linguistics: Industry Track, pages 263–275
January 19–24, 2025. ©2025 Association for Computational Linguistics

lenge. To address these challenges, we introduce Uni-SciSum, a SMSMO framework that leverages the strengths of LLMs while effectively integrating multimodal information within a unified framework. Uni-SciSum employs *BridgeNet*, a Query Transformer (Q-Former) (Li et al., 2023), to fuse different modalities into a fix-length multimodal embedding. It is trained in two stages: first, *modal-to-modal pre-training* aligns different modalities with summaries, extracting modality-specific features relevant for summarization; second, *multimodal instruction tuning* fine-tunes the model for text summary generation and GA selection, learning cross-modal transformations. GA selection is integrated directly into the LLM decoder as an *image token*, extending the textual decoder to handle multimodal outputs. Extensive experiments on two newly introduced SMSMO datasets demonstrate Uni-SciSum's superior performance in generating high-quality summaries, outperforming both uni- and multi-modal models.

2 Related Work

Here we briefly review the literature related to scientific document summarization. We discuss Uni-SciSum relations to multimodal LLMs in Appendix A.

2.1 Multimedia Paper with Summary

Scientific publications are increasingly "multimedia", with publishers like Elsevier and Springer encouraging using GAs, a type of diagrammatic summary or key image, to enhance reading experiences and facilitate searching (Elsevier, 2021; Springer, 2023). The use of GAs is growing rapidly across disciplines, with a 4.5-fold increase of its original level in social science from 2011 to 2015 (Yoon and Chung, 2017) and over 65% of authors in top computer science conferences, such as International Conference on Computer Vision and Conference on Computer Vision and Pattern Recognition, using "teaser figures" (a form of GA) (Yang et al., 2019). Besides images, video is also increasingly used in publication, particularly following COVID-19 when many papers are now presented online. Multimedia papers have been shown to boost publication awareness, with an 8.4-fold increase in retweets and a 2.7-fold increase in paper visits (Ibrahim et al., 2017). To facilitate understanding of multimodal scientific content, it is useful to have an SMSMO system that can generate multimodal summaries from diverse sources, benefiting both editors and readers.

2.2 Scientific Document Summarization

Automatically convert scientific documents into concise summaries has been a classic NLP challenge (Paice, 1980; Teufel and Moens, 2002; Syed et al., 2024). With the increase of multimedia papers, researchers start exploring multimodal summarization. For example, Atri et al. (2021) explored the use of presentation videos for paper abstract generation. Different methods have been proposed to fuse multimodal information, ranging from simple concatenation (Yang et al., 2019) to different optimization strategies, such as contrastive pre-training, Yamamoto et al. (2021). Recent cutting-edge models use transformers to implicitly align data of different modalities (Atri et al., 2023; Kumar et al., 2024). They use cross-modal attention to align individual modalities, but this complex architecture limits its flexibility, making it difficult to adapt to different combinations of input/output data. This work introduces a unified SMSMO framework that utilizes a simple encoder-decoder model to generate summaries from uni- and multi-modal papers. It is trained jointly on data from one/several modalities and handles multimodal output.

3 Model Architecture

As shown in Figure 2, Uni-SciSum comprises several *unimodal encoders* (left), a *BridgeNet* (middle) and a *LLM summary decoder* (right). The encoders process a multimedia paper as input, extracting four feature types: video, audio, text and image. Each modality carries unique features. Inspired by BLIP-2 (Li et al., 2023), we deploy a Q-Former-based BridgeNet to distil multimodal features. It learns to extract a fixed number of modal-specific features from each encoder's outputs using a set of trainable *query vectors* (a.k.a., Q-queries). These queries interact through self- and cross-attention, learning both intra- and inter-modal features relevant to summarization (details in Section 4.1). Since the size of the Q-queries is much smaller than the size of the encoder features, it reduces the computation cost for the decoder. Also, the query size is fixed regardless of the number of modalities, making it more suitable for real-world SMSMO data with variable-length modalities. Finally, we employ Pegasus as the selected LLM for summary generation,

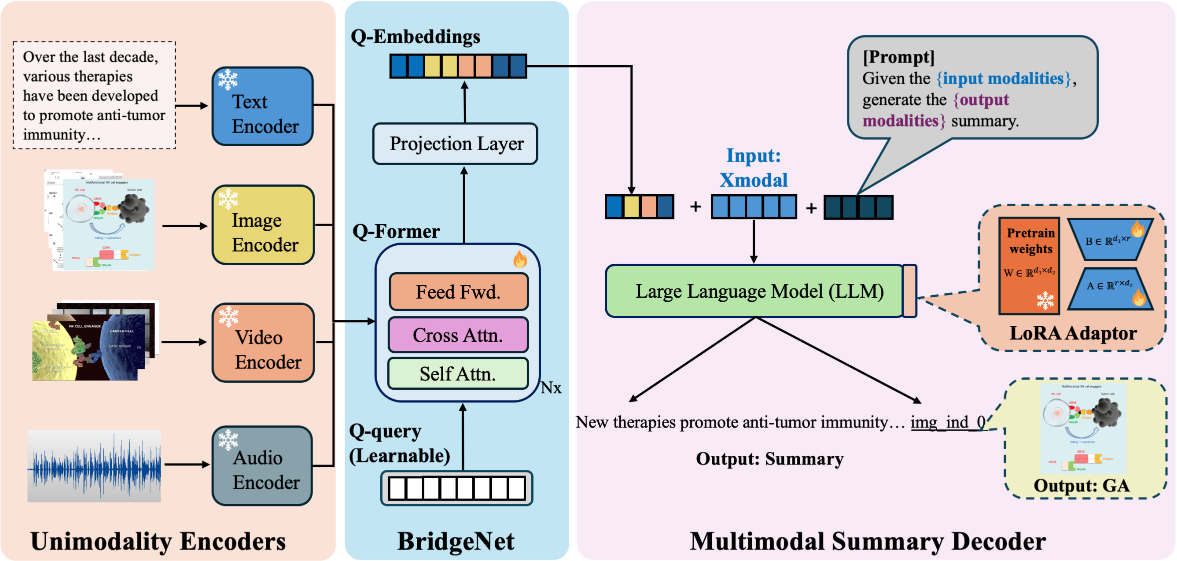


Figure 2: An overview of Uni-SciSum. It connects unimodal encoders to multimodal decoders via BridgeNet. During pretraining, the learnable queries in BridgeNet learn to extract modality-specific features from the encoders. During downstream tasks, the decoder generates embeddings based on different inputs and outputs (guided by the prompt and the learned queries), which the LLM then decodes into the target text summary and GA.

leveraging its exceptional generative performance in many scientific NLP tasks (Zhang et al., 2019). We describe more details in Appendix B.

4 Training Methods

This section describes Uni-SciSum's two-stage training: first, modal-to-modal pre-training aligns different modalities with summaries, enabling the model to learn summary-related multimodal representations. Second, multimodal instruction tuning fine-tunes the model for text summarization and GA selection, facilitating the learning of inter-modal transformations.

4.1 Stage1: Learn Summary-Related Multimodal Representation

Stage 1 focuses on training BridgeNet to effectively connect multimodal features and learn intra- and inter-modality features relevant to summary. This is achieved through two pretraining tasks: *Xmodal-Summary Contrasting* (XSC) and *Xmodal-Summary Matching* (XSM).

Xmodal-Summary Contrasting (XSC). We employ contrastive learning (Radford et al., 2021) to train BridgeNet to extract summary-related features. As illustrated in Figure 3 (left), the q-query and paper summary is fed into BridgeNet to obtain the Xmodal query embeddings and the text embeddings. Here, the self-attention module separately processes the queries and text without any interaction. This enforces the queries to extract intra-modality features specifically from individual encoders, in order to generate representations that align with the corresponding text representations.

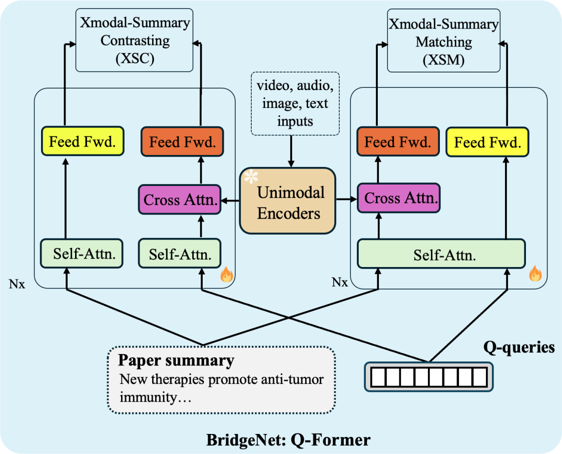


Figure 3: The figure shows BridgeNet's architecture and the two pretraining tasks: XSC (left) and XSM (right). During pretraining, the learnable q-queries interact with each other and various modalities through the self- and cross-attention layers, thereby learning the intra- and inter-modality features relevant for summarization.

Xmodal-Summary Matching (XSM). XSM aims to align cross-modal representations with the text representation. It is a binary classification task, which predicts whether an Xmodal-text pair matches or not (from the same paper). As illustrated in Figure 3 (right), XSM allows the queries and texts to interact through the same self-attention module, thereby allowing the queries to learn finer-grained inter-modality information across Xmodal and texts.

4.2 Stage2: Multimodal Instruction Tuning

Pretraining enables our model to learn summary-related features across different modalities (as captured by the Q-queries). These Q-queries are fed into a multimodal summary generator to produce summary (Figure 2, right). To support transforming information across different modalities, we employ a prompt to guide generation tasks: "*Given* `<input modalities>`*, generate* `<output modalities>` *summary.*", where input modalities can be any combination of video, audio, text, and image; and output modalities include text summaries and/or GAs. GA selection is integrated directly into the decoder using an *index token* appended to the text target (e.g., `img_ind_0` for the first image) (Figure 2, right bottom). This facilitates unified end-to-end training using a Pegasus LLM decoder, eliminating the need for a separate image-scoring module. The prompt and Q-queries are concatenated and fed to the decoder. For training efficiency, we also incorporate Low-Rank Adaptation (LoRA) (Hu et al., 2022) adapters into the LLM. This reduces the trainable parameters of our LLM from 500M to 3M, retaining only 0.6% of the original parameters.

5 Experiment

5.1 Datasets

Due to the lack of multimodal reference in existing scientific summarization datasets (either missing videos or GA), we developed two datasets ($\text{SMSMO}_{\text{mTLDRgen}}$ and $\text{SMSMO}_{\text{Cellpress}}$) to enrich the benchmarks in the SMSMO research area. We use the dataset to pre-train and fine-tune our model.

$\text{SMSMO}_{\text{mTLDRgen}}$ is modified based on the mTLDRgen dataset (Atri et al., 2023), which collected computer science papers to study the effect of multimodal signals (i.e., presentation videos) on text summary generation. Due to the absence of GA targets in the dataset, we employed a heuristic approach to identify key images as proxy labels (details in Appendix C). Briefly, we select images based on a list of summary-related keywords in captions (e.g., "*overall*, *framework*, *overview*, etc."). We compare our list with other keyword filtering and GA selection methods (e.g., ROUGE-ranking, Zhu et al. (2020)). To ensure reliability, two volunteers post-validated the selected images, checking if they represent the paper's abstract. The inter-annotator agreement is 0.72 Cohen's kappa, indicating fair agreement. We obtained 3,224 samples, split into train, validate, and test sets in 8:1:1.

To fine-tune our model for multimodal output generation, we collect papers, video presentations and the corresponding graphical abstract from openly available academic proceedings from the Cell Press[1]. It is a platform where scientists share a short video presentation (with video, text and image) about a paper they have written. The papers are from several virtual conferences, especially in life, physical, earth, and health sciences. We obtained the open PDFs of individual papers and extracted their paragraph text and images (like we did in $\text{SMSMO}_{\text{mTLDRgen}}$). We name this dataset as $\text{SMSMO}_{\text{Cellpress}}$. In total, we collected 190 papers in $\text{SMSMO}_{\text{Cellpress}}$. We divide them into train, valid and test sets in 8:1:1.

5.2 Implementation Detail

Preprocessing. We tokenized all the characters in the source paper text and target summaries with the Longformer's subwords tokenizer (Beltagy et al., 2020).

Model. In the text encoder module of our UniSciSum model, we initialize our embedding matrix using the SciBERT (Beltagy et al., 2019b) model. It contains 30,000 vocabularies with an embedding dimension of 768. The paper text and summaries share the same vocabulary. The paper image feature is extracted by the ResNet-101 encoder (He et al., 2016) and project each image representation to a 768-dimensional vector. We randomly initialize all trainable parameters using a uniform distribution within $[-0.1, 0.1]$.

Training. During training, we configured the model batch size to 2 (due to the restriction of the GPU memory), the learning rate to 0.0001. Additionally, we set the dropout ratio to 0.1. We employ an AdamW (Loshchilov, 2017) optimizer to decouple weight decay from the gradient update and hence prevent overfitting. The experiments are deployed in Pytorch on an NVIDIA GeForce RTX 4090.

Testing. In the testing phase, we configured the decoding beam size as 5. To avoid repetitive tri-

[1] https://www.cell.com/

	Models	Input Modalities				Metrics				
		Text	Image	Video	Audio	$\mathbf{R}_1$	$\mathbf{R}_2$	$\mathbf{R}_L$	$\mathbf{A}_1$	$\mathbf{A}_3$
SSO	LED	✓	-	-	-	8.17	0.37	10.15	-	-
	Long-T5	✓	-	-	-	9.95	0.92	12.57	-	-
	Pegasus	✓	-	-	-	10.56	1.1	11.12	-	-
MSSO	MuLT (Concatenate)	✓	-	✓	✓	11.31	1.99	9.67	-	-
	CFSum	✓	✓	-	-	12.93	0.42	11.29	-	-
	MFN	✓	-	-	✓	11.79	2.1	11.53	-	-
	MAST	✓	-	✓	✓	12.58	2.48	11.4	-	-
MSMO	MSMO	✓	✓	-	-	13.94	0.61	10.46	0.23	0.35
	MLASK	✓	✓	✓	-	14.15	2.87	10.38	0.21	0.32
	Ours	✓	✓	✓	✓	**20.56**	**4.20**	**15.98**	**0.25**	**0.55**

Table 1: Results of our Uni-SciSum and baselines. The top results are **bold**.

grams in the generated summaries, we incorporated trigram blocking (Paulus, 2017).

5.3 Baselines and Evaluation

For evaluation, we compare our model performance against different baselines, covering models of Single Summarization with Single Output (**SSO**), Multi-Modal Summarization with Single-Modal Output (**MSSO**) and Multi-Modal Summarization with Multi-Modal Output (**MSMO**).

Single Summarization with Single Output (SSO). Longformer (LED) (Beltagy et al., 2020) extends the standard seq2seq architecture with sparse attention to handle long text. **Long-T5** (Guo et al., 2022) is the extension of the T5 encoding methods for handling longer input sequences, and **Pegasus** (Zhang et al., 2019) is designed specifically for abstractive summarization for long documents like news and research papers.

Multi-Modal Summarization with Single-Modal Output (MSSO). Multimodal Transformer (MuLT-Concatenate) extends the generic Seq2Seq transformer model. It fuses features of different modalities by concatenating their feature vectors, and the vectors to a transformer decoder to generate textual summaries; **MAST** (Khullar and Arora, 2020) is a multi-modal text summarization model that leverages a trimodal attention mechanism to integrate the text, video and audio modalities at a hierarchical manner, with a first-level pairwise computation of the attention weights between text and other modalities, followed by a second-level attention that focuses on the pairwise attention feature. **MFN** (Liu et al., 2020) is a multistage fusion model that generates summaries based on acoustic and textual input. **CFSum** (Xiao et al., 2023) proposes a contribution network that selects more important parts of images for multimodal summarization and effectively enhances the multimodal representation for summarization.

Multi-Modal Summarization with Multi-Modal Output. MSMO (Zhu et al., 2018) is the first multimodal summarization model with multimodal output, where an attention mechanism is used to fuse the text-image features for better text generation, and the coverage mechanism is used to help select representative images. **MLASK** (Krubiński and Pecina, 2023) develops a Dual-level Interaction Summarizer to generate multimodal summarization based on video and text.

To assess the quality of our generated textual summary, we employ the widely-used ROUGE (Lin, 2004). We follow previous works (Chen et al., 2021; Cohan et al., 2018; Ju et al., 2021) by reporting the F_1 scores of ROUGE-1 (R$_1$), ROUGE-2 (R$_2$) and ROUGE-L (R$_\text{L}$). These scores are computed using the *pyrouge* package[2]. Furthermore, we evaluate the quality of the chosen key image using the top-1 (A$_1$) and top-3 (A$_3$) accuracy metrics introduced by (Yang et al., 2019). These metrics determine whether the positive sample is correctly identified within the top-1 or top-3 positions of the predictions.

6 Results

We evaluate Uni-SciSum against baselines, utilizing SMSMO$_{\text{mTLDRgen}}$ and SMSMO$_{\text{Cellpress}}$ datasets for pre-training and fine-tuning, respectively. Table 1 reports the result on the SMSMO$_{\text{Cellpress}}$ dataset [3]. Overall, Uni-SciSum outperforms other methods in both text summa-

[2] https://github.com/bheinzerling/pyrouge

[3] We also experimented pertaining with SMSMO$_{\text{Cellpress}}$ and fine-tuning it on SMSMO$_{\text{mTLDRgen}}$. The result is reported in Appendix D.1.

Methods		Metrics				
XSC	**XSM**	$\mathbf{R}_1$	$\mathbf{R}_2$	$\mathbf{R}_L$	$\mathbf{A}_1$	$\mathbf{A}_3$
X	X	14.18	1.86	10.90	0.13	0.43
X	✓	15.60	1.85	11.70	0.15	0.45
✓	X	17.59	2.25	13.79	0.24	0.53
✓	✓	**20.56**	**4.20**	**15.98**	**0.25**	**0.55**

Table 2: Results on the effect of different methods of pre-training BridgeNet. The top results are **bold**.

Models	Metrics				
	$\mathbf{R}_1$	$\mathbf{R}_2$	$\mathbf{R}_L$	$\mathbf{A}_1$	$\mathbf{A}_3$
Ours_{text}	12.15	1.18	8.81	0.05	0.15
$\text{Ours}_{text+video}$	14.20	1.10	10.71	0.15	0.4
$\text{Ours}_{text+video+audio}$	16.08	1.88	12.46	0.15	0.4
Ours_{all}	**20.56**	**4.20**	**15.98**	**0.25**	**0.55**

Table 3: Experiment results on the ablation study on different modalities. The top results are **bold**.

rization and GA selection. Compared to unimodal SSO methods, Uni-SciSum shows better performance in text summarization, highlighting its advantages of using multimodal data. Moreover, Uni-SciSum outperforms multimodal methods (both MSSO and MSMO), demonstrating the effectiveness of leveraging cross-modal salient information for the summarization process. The results show that Uni-SciSum can distil knowledge from unimodal encoders pre-trained on large-scale datasets. Particularly, our BridgeNet effectively exploits the modality-specific knowledge embedded in different pre-trained models to perform text summarization, and adapt it across related task of GA selection. Through XSC and XSM pre-training, the model's query representations acquire comprehensive summary-related information within and across modalities, effectively generating text summaries and identifying target images. Given the shared features of summary-related signals and our multimodal prompt tuning, adapting Uni-SciSum to other new tasks (e.g., video→text) also becomes easier (as later shown in Table 3).

6.1 Ablation Study

Ablating Pre-training. Table 2 demonstrates the impact of pre-training on BridgeNet performance. It helps BridgeNet learn relevant multimodal features, thereby reducing the burden on the LLM and leading to the best summary score (shown at the bottom of the table). Conversely, removing either XSC or XSM results in lower scores, indicating the importance of both intra- and inter-modality pre-training for effective multimodal summarization.

Ablating Modalities. Table 3 shows the models' performance when we fine-tune Uni-SciSum on different modalities (text, video, audio and/or image). We observe that combining all modalities leads to improved performance in both text and image tasks, demonstrating Uni-SciSum's effectiveness in leveraging multiple modalities for enhanced cross-modal feature extraction and improved multimodal summarization. We provide full ablation studies on different modality combinations in Appendix D.2.

Modules		Metrics				
BridgeNet	**LLM**	$\mathbf{R}_1$	$\mathbf{R}_2$	$\mathbf{R}_L$	$\mathbf{A}_1$	$\mathbf{A}_3$
Q-Former	Pegasus	**20.56**	**4.20**	**15.98**	**0.25**	**0.55**
Q-Former	LED	18.48	3.76	11.73	0.15	0.50
Q-Former	Long-T5	16.13	3.29	13.46	0.20	0.50
Linear	Pegasus	12.05	1.61	8.05	0.15	0.45
Linear	LED	12.37	1.24	8.83	0.15	0.45
Linear	Long-T5	11.35	0.93	9.60	0.15	0.45

Table 4: Results on ablating different querying methods and decoder LLMs. The top results are **bold**.

Ablating Query Methods and Decoders. Table 4 shows that replacing the Q-Former in BridgeNet with a linear layer worsens summary generation, resulting in an average decrease of 45.3% and 17.2% in text and image scores, respectively. Also, replacing the Pegasus LLM decoder with Longformer or Long-T5 decreases performance. These findings demonstrate Q-Former's effectiveness in extracting summary-related information from multimodal data and Pegasus's strength in text generation. Table 5 further analyzes Pegasus' performance when pre-trained on different text genres, including social media (Pegasus_{reddit}), news (Pegasus_{xsum}), papers (Pegasus_{arxiv}) and a mix (Pegasus_{large}). The best results came from a PubMed-trained Pegasus model, demonstrating the importance of domain-specific LLM for scientific NLP.

Models	Metrics				
	$\mathbf{R}_1$	$\mathbf{R}_2$	$\mathbf{R}_l$	$\mathbf{A}_1$	$\mathbf{A}_3$
Pegasus_{reddit}	15.70	1.04	11.48	0.15	0.40
Pegasus_{xsum}	14.83	1.94	11.72	0.20	0.45
Pegasus_{arxiv}	16.59	3.75	10.39	0.15	0.40
Pegasus_{large}	17.48	3.84	14.15	0.15	0.40
Pegasus_{pumbed}	**20.56**	**4.20**	**15.98**	**0.25**	**0.55**

Table 5: Results on ablating LLM pre-trained on different document genres. The top results are **bold**.

Reference summary: PAM dopaminergic neurons are active during flight and require octopaminergic inputs. Flight-regulating PAM neurons project to the β'1 lobe of the mushroom body. Shorter flight bouts are observed upon activation of GABAergic β'1 output neurons. PAM neurons inhibit GABAergic β'1 output neurons to support extended flight bouts.

Pegasus: flight is a complex behavior that requires the integration of multiple sensory inputs with flight motor output . previous genetic studies identified central brain monoaminergic neurons that modulate central brain monoaminergic and octopaminergic neurons that modulate sustained flight bouts to higher classes of flight- and mechanosensory neurons that project to the mushroom body brain .

CFSum: Insect flight is a complex behavior that requires the integration of multiple sensory inputs with flight motor output. Although previous genetic studies identified central brain monoaminergic neurons that modulate Drosophila flight, neuro-modulatory circuits underlying sustained flight are not identified.

MLASK: As in the early alignment, conservation of their primary sequences, biological tissues, synthesis, and Fec-based critical capacity, vaccination, pylation, IrRNA-associated infections, and evolution.

Ours: PDP Sparrow flight-based flight trains underlying flight neuronal circuits. The flight amplitudes and function during flight are reduced in the absence of dopamine control. The perturbations influence flight mechanism in the mushroom brain. The transient flight mechanism is under dynomps control of the flight" assumed assumed.

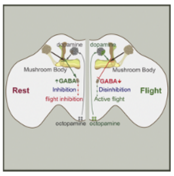
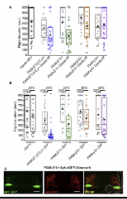
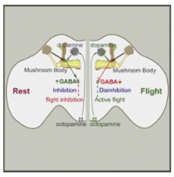
Ground Truth	***MLASK***	***Ours***
		

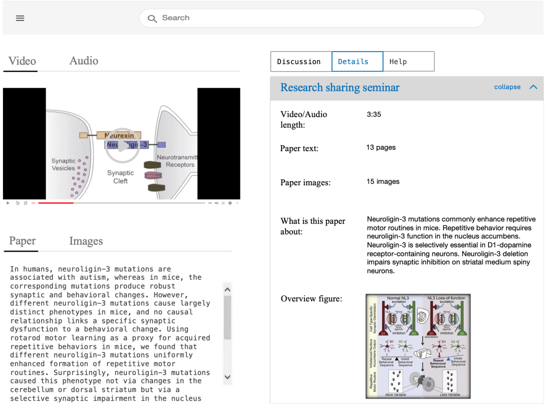
Table 6: Illustration of the generated summary from baselines and Uni-SciSum.

7 Case Study

Table 6 compares the summary outputs by the best-performing models in the SSO (Pegasus), MSSO (CFSum), and MSSO (MLASK) categories. We also include the abstract for reference (Table 6, top). Here, we observe that our Uni-SciSum offers finer-grained information compared to others. For example, it identifies details relating to the role of dopamine in regulating flight behaviour and the underlying neuronal circuits, offering a more nuanced understanding of the flight mechanism. Conversely, CFSum and Pegasus capture general aspects of the flight process. Meanwhile, MLASK struggles to capture relevant flight-related information, focusing instead on unrelated biological aspects, such as tissue synthesis and evolution, without addressing the key neural mechanisms involved in flight.

8 Conclusion

To address the growing need for effective multimodal processing in scientific NLP, this work introduces Uni-SciSum, a unified SMSMO architecture designed to generate multimodal summaries from multimedia papers. Uni-SciSum's design comprises a Q-Former-based BridgeNet for effective multimodal representation fusion; a two-stage training strategy consisting of modal-to-modal pre-training and cross-modal instruction tuning to ensure alignment and adaptation across modalities/tasks; and a specialized LLM decoder that can generate both text and image tokens, thereby eliminating the need for a separate image scoring module. Experiments show that our model improves the quality of multimodal output on both real human-labeled and automatically constructed datasets, outperforming both uni- and multi-modality models. This work contributes to the advancement of scientific communication by introducing a new framework (with data and models) for efficient summarization of complex multimedia research. We plan to deploy Uni-SciSum on an AI platform (Figure 4), initially for research seminar summarization on campus, and subsequently exploring its integration with other AI tools/tasks (e.g., paper video question-answering) to facilitate the dissemination of educational resources for remote learning.



Research sharing seminar

Video/Audio length: 3:35

Paper text: 13 pages

Paper images: 15 images

What is this paper about: Neuroligin-3 mutations commonly enhance repetitive motor routines in mice. Repetitive behavior requires neuroligin-3 function in the nucleus accumbens. Neuroligin-3 is selectively essential in D1-dopamine receptor-containing neurons. Neuroligin-3 deletion impairs synaptic inhibition on striatal medium spiny neurons.

Overview figure:

```
In humans, neuroligin-3 mutations are
associated with autism, whereas in mice, the
corresponding mutations produce robust
synaptic and behavioral changes. However,
different neuroligin-3 mutations cause largely
distinct phenotypes in mice, and no causal
relationship links a specific synaptic
dysfunction to a behavioral change. Using
rotarod motor learning as a proxy for acquired
repetitive behaviors in mice, we found that
different neuroligin-3 mutations uniformly
enhanced formation of repetitive motor
routines. Surprisingly, neuroligin-3 mutations
caused this phenotype not via changes in the
cerebellum or dorsal striatum but via a
selective synaptic impairment in the nucleus
```

Figure 4: Proposed deployment of Uni-SciSum within the AI platform.

Acknowledgments

The work is supported by the Hong Kong RGC ECS (LU23200223/130393) and Internal Grants of Lingnan University, Hong Kong (code: LWP20018/871232, DR23A9/101194, DB23B5/102083, DB23AI/102070 and 102241).

References

Yash Kumar Atri, Vikram Goyal, and Tanmoy Chakraborty. 2023. Fusing multimodal signals on hyper-complex space for extreme abstractive text summarization (tl; dr) of scientific contents. In *Proceedings of the 29th ACM SIGKDD Conference on*

Knowledge Discovery and Data Mining, pages 3724–3736.

Yash Kumar Atri, Shraman Pramanick, Vikram Goyal, and Tanmoy Chakraborty. 2021. See, hear, read: Leveraging multimodality with guided attention for abstractive text summarization. *Knowledge-Based Systems*, 227:107152.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019a. Scibert: A pretrained language model for scientific text. 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 3615–3620.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019b. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Xiuying Chen, Hind Alamro, Mingzhe Li, Shen Gao, Xiangliang Zhang, Dongyan Zhao, and Rui Yan. 2021. Capturing relations between scientific papers: An abstractive model for related work section generation. Association for Computational Linguistics.

Christopher Clark and Santosh Divvala. 2016. Pdffigures 2.0: Mining figures from research papers. In *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, pages 143–152.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of NAACL-HLT*, pages 615–621.

Elsevier. 2021. How to produce a good visual abstract, tools and resources for authors. Https://www.elsevier.com/authors/tools-and-resources/visual-abstract.

Grobid. 2020. Grobid parser. `https://github.com/kermitt2/grobid`.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. Longt5: Efficient text-to-text transformer for long sequences. *Findings of the Association for Computational Linguistics: NAACL 2022*.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

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

Andrew M Ibrahim, Keith D Lillemoe, Mary E Klingensmith, and Justin B Dimick. 2017. Visual abstracts to disseminate research on social media: a prospective, case-control crossover study. *Annals of surgery*, 266(6):e46–e48.

Jiaxin Ju, Ming Liu, Huan Yee Koh, Yuan Jin, Lan Du, and Shirui Pan. 2021. Leveraging information bottleneck for scientific document summarization. *arXiv preprint arXiv:2110.01280*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota.

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

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

Sandeep Kumar, Guneet Singh Kohli, Tirthankar Ghosal, and Asif Ekbal. 2024. Longform multimodal lay summarization of scientific papers: Towards automatically generating science blogs from research articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10790–10801.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

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

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.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Chris D Paice. 1980. The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. In *Proceedings of the 3rd annual ACM conference on Research and development in information retrieval*, pages 172–191.

R Paulus. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

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

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, pages 1–20.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2024. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36.

Springer. 2023. What is a graphical abstract and why do i need one for my paper? Https://solutions.springernature.com/blogs/visibility/what-is-a-graphical-abstract-and-why-do-i-need-one-for-my-paper.

Shahbaz Syed, Khalid Al Khatib, and Martin Potthast. 2024. Tl; dr progress: Multi-faceted literature exploration in text summarization. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 195–206.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*, 28(4):409–445.

Min Xiao, Junnan Zhu, Haitao Lin, Yu Zhou, and Chengqing Zong. 2023. Cfsum: Coarse-to-fine contribution network for multimodal summarization. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Shintaro Yamamoto, Anne Lauscher, Simone Paolo Ponzetto, Goran Glavaš, and Shigeo Morishima. 2021. Visual summary identification from scientific publications via self-supervised learning. *Frontiers in Research Metrics and Analytics*, 6:719004.

Sean T Yang, Po-Shen Lee, Lia Kazakova, Abhishek Joshi, Bum Mook Oh, Jevin D West, and Bill Howe. 2019. Identifying the central figure of a scientific paper. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1063–1070. IEEE.

JungWon Yoon and EunKyung Chung. 2017. An investigation on graphical abstracts use in scholarly articles. *International Journal of Information Management*, 37(1):1371–1379.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553.

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.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *Preprint*, arXiv:1912.08777.

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.

A Related Works

Section 2 in our main paper reviewed the literature in scientific summarization. Here, we describe **Multimodal Large Language Models**.

A.1 Multimodal Large Language Models

Large Language Models (LLMs) like BERT (Kenton and Toutanova, 2019) and the GPT (Brown et al., 2020) family have received more attention due to their performance and potential applications. Some variants like SciBERT (Beltagy et al., 2019a), Longformer (Beltagy et al., 2020) and Long-T5 (Guo et al., 2022) have been adapted for textual NLP tasks within the scientific domain. Recent research has focused on extending LLMs to multimodal interactions, encompassing video, audio, image, and text modalities. Two primary approaches have emerged. The first approach positions LLMs as a multitask processor, mapping different modal tasks to a unified space. For example, BLIP-2 (Li et al., 2023) maps images to text space using Q-Former, while Video-LLaMA (Zhang et al., 2023) maps audio and vision modalities via Q-Former. The second approach uses LLMs as a task coordinator, connecting them to specialized downstream models. For example, Shen et al. (2024) build the HuggingGPT framework. It uses GPT to conduct task planning when receive a user request, select models according to their function descriptions available in Hugging Face, and execute each subtask with the dedicated model.

Current multimodal LLM approaches, while promising, often lack the flexibility to handle diverse modality combinations. They are either limited to specific pairings (e.g., image-text in Q-Former) or require modality-specific modules (e.g., HuggingGPT). Our work offers a more streamlined and adaptable solution that enhances flexibility and simplifies the architecture. Particularly, our work extends Q-Former to incorporate four modalities (video, audio, text, image) and introduces `index tokens` formulation for direct image selection, eliminating the need for a separate scoring module. This unified framework enables a single LLM decoder to process both uni- and multi-modalities data, providing a more general and efficient approach to SMSMO tasks.

B Model Architecture

Section 3 in our main paper mentions our model architecture. Here, we provide the details of our **encoders** and **BridgeNet**:

B.1 Multimodal Encoders

We use the following four feature encoders corresponding to the input modalities used in SMSMO:

- **Text**: To encode the paper text feature, we utilized the SciBERT (Beltagy et al., 2019b) model, specifically designed to handle the complexities and nuances inherent in scientific texts.
- **Image**: We the ResNet (He et al., 2016) model to handle the image features (e.g., figures, tables, and algorithms) in the scientific paper.
- **Video**: We use a 2048-dimensional feature vector per group of 16 frames, which is extracted from the videos using a ResNeXt-101 3D CNN trained to recognize 400 different actions (Hara et al., 2018). This results in a sequence of feature vectors per video.
- **Audio**: We use the concatenation of 40- dimensional Kaldi (Povey et al., 2011) filter bank features from 16kHz raw audio using a time window of 25ms with 10ms frame shift and the 3-dimensional pitch features extracted from the dataset to obtain the final sequence of 43-dimensional audio features.

B.2 BridgeNet

Inspired by BLIP-2 (Li et al., 2023), we employ a Q-Former-based BridgeNet. It summarizes the variable-length embeddings from each encoder's outputs within a given number of learnable query extracts a fixed number of modal-specific features from each encoder's outputs using a set of trainable *query vectors* (a.k.a., Q-queries). The queries interact with each other through self-attention layers, and interact with the frozen encoders' features through cross-attention layers. Since the size of the Q-queries is much smaller than the size of the encoder features, it significantly reduces the computation cost for the decoder.

Formally, let X_m be the m-th modality features extracted from its corresponding unimodal encoder (referred to as Xmodal features henceforth). Q-queries is a set of learnable vector denoted as $q \in \mathbb{R}^{n_q \times d_q}$, where n_q and d_q represent the number

	$\text{SMSMO}_{\text{mTLDRgen}}$			$\text{SMSMO}_{\text{Cellpress}}$		
	Train	Valid	Test	Train	Valid	Test
Num of docs	2,583	320	321	150	20	20
Avg. img num	7.07	6.62	6.88	8.12	6.91	7.11
Avg. sent num	222.14	223.15	221.21	232.12	267.12	237.31
Avg. video/audio len (s)	744.11	717.12	728.21	274.51	290.21	315.12

Table 7: Corpus statistics of our dataset.

Paper Type	**Size**	**Avg. sent num**	**Avg. img num**	**Avg. video/audio len (s)**
ACL	1,174	218	8	1,031
CVPR	301	226	10	384
ICCV	69	227	10	401
ICML	687	256	6	799
IJCAI	919	205	7	489
NeurIPS	74	209	5	454

Table 8: Data Source Distribution on the $\text{SMSMO}_{\text{mTLDRgen}}$ dataset.

Keywords
flow chart, flowchart, illustration, general block diagram, system structure, system architecture, overall, overview, framework, workflow, structure, flow, demonstration, graphic visualization, graphical (model), theoretical model

Table 9: The keywords we use to identify the key figures (i.e., GA) in our $\text{SMSMO}_{\text{mTLDRgen}}$ dataset. The key image of individual papers is determined by the number of keywords each image caption contains. If there is a tie, the image that appears earlier in the paper will be taken. Images which can not align with any keywords are excluded.

and dimension of query vector. First, we input the Q-queries into the self-attention mechanism:

$$A^{self} = \text{softmax}\left(\frac{qW_q^{self}(qW_k^{self})^T}{\sqrt{d_k}}\right) qW_v^{self}, \tag{1}$$

where $W_q^{self} \in \mathbb{R}^{d_q \times d_k}$, $W_k^{self} \in \mathbb{R}^{d_q \times d_k}$, and $W_v \in \mathbb{R}^{d_q \times d_v}$ are the learnable weight matrices for queries, keys, and values (resp.). And d_k represents the dimensions of the keys. The output $A^{self} \in \mathbb{R}^{n_q \times d_v}$ is then used for the cross-attention mechanism with the Xmodal feature X:

$$A_x^{cross} = \text{softmax}\left(\frac{A^{self}W_q(XW_k)^T}{\sqrt{d_k}}\right) XW_v, \tag{2}$$

where $A_x^{cross} \in \mathbb{R}^{n_x \times n_q \times d_v}$ represents the cross-attention output. The matrices $W_q \in \mathbb{R}^{d_v \times d_k}$, $W_k \in \mathbb{R}^{d_x \times d_k}$, and $W_v \in \mathbb{R}^{d_x \times d_v}$ are the learnable weight matrices for queries, keys, and values. After the feed-forward layer, the final embedding of Q-queries of Xmodal is denoted as $M_{qx} \in \mathbb{R}^{n_x \times n_q \times d_q}$. It represented the modal-specific feature relevant to summarization, as distilled from individual unimodal encoders.

Q-Former's weights are initialized from SciBERT, a BERT LLM pretrained on scientific publications, which has shown promising performances in many scientific NLP tasks (Beltagy et al., 2019b). The cross-attention module is added into the Q-Former every two layers and is randomly initialized.

C Dataset Construction

We created two datasets: $\text{SMSMO}_{\text{mTLDRgen}}$ and $\text{SMSMO}_{\text{Cellpress}}$. Their statistics are presented in Table 7.

$\text{SMSMO}_{\text{mTLDRgen}}$ is a modified version based on the mTLDRgen dataset (Atri et al., 2023), which collected conference papers in computer science to study the effect of multimodal signals (i.e., presentation videos) on text summary generation. In mTLDRgen, the authors collected the presentation videos from well-known conferences in computer science (e.g., ACL, ICCV, CVPR, etc., see Table 8); and used them to generate the corresponding human-written summary (TLDR). Here, we utilize the paper sources from mTLDRgen to build our new dataset. Particularly, we obtained the PDFs of individual papers in $\text{SMSMO}_{\text{mTLDRgen}}$, and extracted their body text and images using Grobid (Grobid, 2020) and Pdffigures (Clark and Divvala, 2016) (resp.). We filter out the data examples which contain no images. We take the paper abstracts as the target summary for geneartion since we cannot obtain the TLDR summary from the authors. Then, we employ a heuristic method to generate the pseudo image selection labels for our data. Specifically, in research articles, images that provide summary information are often captioned with keywords like "overall, framework, overview, etc." (see Table 9). Here, we leverage this property and use a list of summary-related keywords to identify the key images for individual papers. We didn't prioritize the keywords, and we picked the image with the caption that contains most of the keywords (In case there is a tie, we picked the larger image). We compare our keyword lists with the ones generated automatically by Rapid Automatic Keyword Extraction (RAKE) (Rose et al., 2010), TextRank (Mihalcea and Tarau, 2004). We also compare our methods with Order-ranking and ROUGE-ranking proposed by (Zhu et al., 2020), which extract GA by considering the image's order appearing in the paper and the ROUGE value between individual image captions and the text abstract. For comparison, we manually labelled 100 key figures in $\text{SMSMO}_{\text{mTLDRgen}}$. We compare this ground truth with the results obtained from ours and other methods, achieving a top-3 accuracy of 62%, notably higher than the one obtained from the RAKE (53%), TextRank (51%), Order-ranking (47%) and ROUGE-ranking (58%). Consequently, we use our keyword list to obtain the key figure in $\text{SMSMO}_{\text{mTLDRgen}}$. To ensure the test set is reliable, two volunteers are engaged for post-validation, in which they check if the selected figures can represent the paper given its abstract. The inter-annotator agreement amounts to 0.72 Cohen's kappa, which denotes a fair agreement. Using our methods, we get 3,224 data samples. We divide them into train, valid and test sets following the ratio in (Atri et al., 2021) (8:1:1).

We also create $\text{SMSMO}_{\text{Cellpress}}$, an SMSMO dataset with gold GA labels. Particularly, we collect papers, video presentations and the corresponding graphical abstract from openly available academic proceedings from the Cell Press[4]. It is a platform where scientists share a short video presentation (with video, text and image) about a paper they have written. The papers are from several virtual conferences, including life, physical, earth, and health sciences. We obtained the open PDFs of individual papers and extracted their paragraph text and images (like we did in $\text{SMSMO}_{\text{mTLDRgen}}$). In total, we collected 190 papers in $\text{SMSMO}_{\text{Cellpress}}$. We divide them into train, valid and test sets in 8:1:1.

D More Experiment Results

Section 6 in our main paper mentions the main results. Here, we provide further results on other datasets (**$\text{SMSMO}_{\text{mTLDRgen}}$**) and **modalities**:

D.1 Results on $\text{SMSMO}_{\text{mTLDRgen}}$ Dataset

In this part, we pre-train our Uni-SciSum using the $\text{SMSMO}_{\text{Cellpress}}$ dataset, followed by fine-tuning and testing it on $\text{SMSMO}_{\text{mTLDRgen}}$. Table 10 shows the results. We can see that our Uni-SciSum outperforms other models in both text summary generation and GA selection. Despite being pre-trained on a small dataset of 190 samples ($\text{SMSMO}_{\text{Cellpress}}$), our model is still able to demonstrate its ability to acquire cross-modal knowledge during the pre-training phase and subsequently apply it during the fine-tuning steps.

D.2 Ablating Modalities

Table 11 presents the complete results of our modality ablation study, as described in Section 6.1-Table 3. The result demonstrates that incorporating multimodal information essentially improves summarization performance compared to using text alone (the top row). Specifically, combining text with visual modalities (video and/or image) yields better results than using text and audio. This highlights the importance of visual data for summarization. Furthermore, the best performance is achieved when integrating text, video, and audio, suggesting a synergistic effect between these modalities.

[4] https://www.cell.com/

	Models		Input Modalities					Metrics		
		Text	Image	Video	Audio	$\mathbf{R}_1$	$\mathbf{R}_2$	$\mathbf{R}_L$	$\mathbf{A}_1$	$\mathbf{A}_3$
SSO	LED	✓	-	-	-	12.2	4.48	14.73	-	-
	Long-T5	✓	-	-	-	10.36	3.75	13.16	-	-
	Pegasus	✓	-	-	-	20.37	6.41	18.95	-	-
MSSO	MuLT (Concatenate)	✓	-	✓	✓	19.79	5.1	10.53	-	-
	MFN	✓	-	-	✓	25.15	6.95	13.10	-	-
	MAST	✓	-	✓	✓	26.20	7.08	13.13	-	-
	CFSum	✓	✓	-	-	24.31	7.99	11.67	-	-
MSMO	MSMO	✓	✓	-	-	27.84	8.68	15.52	0.26	0.53
	MLASK	✓	✓	✓	-	28.32	8.31	13.57	0.23	0.53
	Ours	✓	✓	✓	✓	**42.22**	**13.14**	**22.88**	**0.27**	**0.58**

Table 10: Results on the SMSMO$_{\text{mTLDRgen}}$ dataset, comparing the performance of our Uni-SciSum model against various baselines across across Single Summarization with Single Output (SSO), Multi-Modal Summarization with Single-Modal Output (MSSO) and Multi-Modal Summarization with Multi-Modal Output (MSMO). The top results are **bold**.

Models			**Metrics**		
	$\mathbf{R}_1$	$\mathbf{R}_2$	$\mathbf{R}_L$	$\mathbf{A}_1$	$\mathbf{A}_3$
Uni-SciSum$_{text}$	12.15	1.18	8.81	0.05	0.15
Uni-SciSum$_{image}$	8.81	0.69	6.34	0.15	0.2
Uni-SciSum$_{video}$	9.74	0.58	9.31	0.05	0.15
Uni-SciSum$_{audio}$	6.36	0.58	6.19	0.05	0.1
Uni-SciSum$_{text+video}$	14.20	1.10	10.71	0.15	0.4
Uni-SciSum$_{text+audio}$	13.16	1.68	9.23	0.1	0.4
Uni-SciSum$_{text+image}$	14.07	1.32	11.10	0.15	0.4
Uni-SciSum$_{text+video+audio}$	16.08	1.88	12.46	0.15	0.4
Uni-SciSum$_{text+video+image}$	16.84	2.15	13.58	0.2	0.45
Uni-SciSum$_{text+audio+image}$	16.44	1.82	12.71	0.15	0.45
Uni-SciSum$_{ours}$	**20.56**	**4.20**	**15.98**	**0.25**	**0.55**

Table 11: Ablation study on different modalities on the SMSMO$_{\text{Cellpress}}$. The top results are **bold**.