Progressive Visual Refinement for Multi-modal Summarization

Anonymous ACL submission

Abstract

Multi-modal summarization (MMS) has emerged as a critical research area driven by the proliferation of multimedia content, focusing on generating condensed summaries by cross-modal complementary information synthesis. Previous studies have demonstrated the effectiveness of heterogeneous fusion paradigms, particularly through visual-centric feature extraction mechanisms, in constructing cross-modal representations that yield substantial performance gains. However, the use of multi-modal information and the inter-correlation among textual content, visual elements, and summary generation have been still insufficiently explored. We propose the Patch-Refined Visual Information Network (PRVIN) to address the insufficient exploitation of visual information. The essential patch selector and patch refiner components in PRVIN work collaboratively to progressively identify and refine critical visual features. An additional vision-to-summary alignment mechanism is also introduced to enhance the semantic connections between multi-modal representations and summary outputs. Extensive experiments conducted on two public MMS benchmark datasets demonstrate the superiority of PRVIN while quantitatively validating the crucial role of comprehensive visual information utilization in MMS tasks.[1]

1 Introduction

With the exponential growth of multimedia content, e.g., news articles with images, instructional videos with audio narration, and social media posts combining texts and visuals, the need to efficiently process and distill information across modalities has become critical. Multi-modal summarization (MMS) addresses this challenge by generating concise and coherent summaries that integrate key information from heterogeneous sources such as text, image, audio, and video (Li et al., 2018; Sanabria et al., 2018; Zhu et al., 2018; Jangra et al., 2020; Palaskar et al., 2019). Unlike traditional text-only summarization, MMS requires models to not only understand intra-modal relationships but also capture cross-modal interactions to identify salient content and synthesize unified outputs.

Previous studies have focused on effectively extracting visual information and combining it with textual information before injecting it into the summarization model. For instance, Yu et al. (2021) injected visual information into pre-trained language models (PLMs) by designing an attention-based add-on layer. Liu et al. (2020) proposed a multi-stage fusion network with a fusion forget gate that models fine-grained cross-modal interactions. Liang et al. (2023) devised two auxiliary tasks including a vision-to-summary task and a masked image modeling task to enhance visual understanding. Zhang et al. (2024) extended BART by integrating a dual weight-sharing multi-modal encoder that concurrently processes textual and visual data alongside entity-specific visual information and introduced a gating mechanism to effectively utilize the resulting multi-modal information for text generation. Nonetheless, these methods still under-utilize informative visual cues critical for summarization while allowing for extraneous visual data, which impair performance. These risks will lead to the visual inputs failing to provide effective information for the summarization process and over-reliance on textual sources, thereby limiting the model's multi-modal capabilities.

To alleviate the above mentioned issues, we propose a novel Patch-Refined Visual Information Network (PRVIN), which uses a visual patch as the smallest unit for selecting and refining visual information. To address the challenge of incomplete utilization of visual information, PRVIN introduces an innovative two-stage framework that systematically optimizes visual patch selection and refine-

[1] Our code will be available at https://github.com/

ment. In the first stage, the proposed essential patch selector uses a dual-alignment mechanism to evaluate the relevance of each visual patch to both input text and task objectives. This module generates alignment scores through multi-modal correlation analysis, enabling selective retention of the most pertinent patch tokens on the basis of a predetermined threshold ratio. The subsequent stage features a novel patch refinement network (PRN) that implements hierarchical visual information processing. Building upon the previously selected patches, the PRN executes secondary filtering to obtain two complementary patch token sequences. These sequences undergo iterative refinement through cross-attention operations, allowing for dynamic information exchange and feature enhancement at the patch level. This cascaded refinement architecture effectively enhances visual information completeness while maintaining computational efficiency through progressive token reduction. We also introduce a vision-centric task that exclusively uses visual inputs through masking text inputs. This auxiliary task enhances cross-modal alignment by establishing direct semantic connections between visual patterns and summary generation, while simultaneously improving visual-representation learning through modality-specific constraints, and mitigates the text preference bias commonly observed in MMS, where models tend to over-rely on textual cues while under-utilizing visual information. Finally, we conducted extensive experiments on two public MMS datasets, and the experimental results indicate the effectiveness of our PRVIN method.

2 Methodology

2.1 Problem Formulation

Given a source text D and its corresponding image V, where $D = (t_1, t_2, ..., t_m)$ is a sequence of m tokens in the source text, the objective is to generate a brief summary $S = \{s_1, s_2, ..., s_l\}$ that effectively captures the essential information from both modalities. The model learns a mapping function $f : (D, V) \rightarrow S$.

2.2 Text Encoding

We employ a pre-trained network, e.g., BART-base (Lewis et al., 2020), to encode the source text, prepending a [CLS] token to capture global semantic information. The resulting text representation is formulated as $T = (t_{cls}, t_1, t_2, ..., t_m)$, where t_{cls} serves as a condensed semantic representation of the entire text sequence.

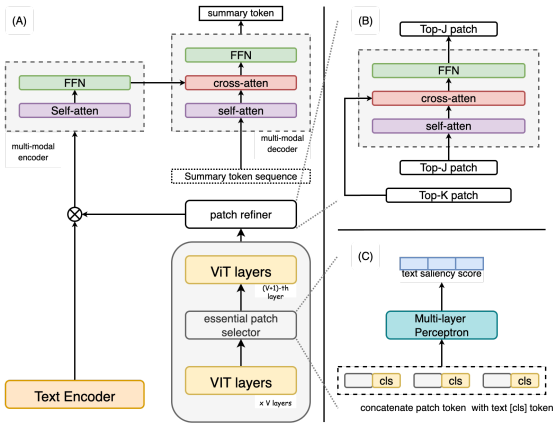

Figure 1: (a) Overview of PRVIN. The two networks in right constitute integral components of our progressive patch refinement process, i.e., (b) the patch refiner and (c) the essential patch selector.

2.3 Visual Encoding

The image encoding process employs the vision Transformer (ViT) to map high-dimensional pixel data into compact latent representations. We divide each image into n non-overlapping patches and prepend a [CLS] token to the patch sequence to represent the entire image: $V = (p_{cls}, p_1, p_2, ..., p_n)$. Each patch embedding p_i encodes local visual information, while p_{cls} aggregates global image semantics through transformer layers.

2.4 Essential Patch Selector

We propose an essential patch selector to determine the importance of each patch and its relevance to the summary and source text. Drawing inspiration from conventional extractive summarization methods (Zhou et al., 2018; Liu and Lapata, 2019), we formulate patch selection as a classification problem, aiming to determine whether a patch aligns with the summary's semantics and should be selected. As illustrated in Fig 1, the selector is integrated between the $\mathcal{V}$-th and ($\mathcal{V}$+1)-th layers of the ViT in the image encoder. To establish an explicit correspondence between each patch and the source text, we concatenate each individual visual patch token with the text [CLS] token: $\bar{p}_i = concat(p_i^k, t_{cls})$, where p_i^k represents the i-th patch in the output of the $\mathcal{V}$-th ViT layer.

The concatenated embeddings are then fed to a multi-layer perceptron (MLP) which is used to predict the relevance between patches and text in-

formation (both the source and summary). The output of the MLP is passed through a sigmoid-activation function to obtain the relevance score r_i: $r_i = Sigmod(MLP(\bar{p}_i))$. We then extract top-$k$ patches in the patch token sequence based on their relevance score r, and the new patch sequence will be fed to the ($\mathcal{V}$+1)-th ViT layer: $\hat{V} = (\hat{p_{cls}}, \hat{p_1}, \hat{p_2}, ..., \hat{p_k})$. This selection process reduces computational complexity while preserving semantically salient visual information.

2.5 Patch Refinement

While we initially select significant patches based on the relevance scores, we posit that further refinement could reduce the length of visual sequences and enhance the training and inference efficiency of the vision-encoder network. Thus, we introduce a post-processing refinement layer superimposed on the base vision encoder, implementing a multi-stage feature enhancement strategy to actualize patch refinement. Specifically, we first select top-j patches ($j < k$) based on the relevance score in the previous step: $\dot{V} = (\dot{p_{cls}}, \dot{p_1}, \dot{p_2}, ..., \dot{p_j})$.

The patch refiner consists of several transformer-decoder blocks. Patch sequence $\dot{V}$ then undergoes processing through the self-attention layer, followed by the cross-attention operation with $\hat{V}$: $\dot{V} = CrossAttn(SelfAttn(\dot{p_{cls}}, \dot{p_1}, ..., \dot{p_j}), \hat{V})$.

Consequently, the final representation for patch sequence $\dot{V}$ ensures the refinement of the visual sequence while preserving the structural and visual integrity of each patch.

2.6 Multi-modal Decoder

To fuse textual and visual modalities, the outputs of the patch refiner and the text encoder are concatenated and fed into the multi-modal encoder to obtain a cross-modal representation: $C = concat(T, \dot{V})$. Then, the cross-modal representation is fed into a multi-modal decoder composed of transformer decoder blocks to generate the corresponding summary: $\hat{s_k} = Decoder(C, \hat{s_1}, \hat{s_2}, ..., \hat{s}_{k-1})$, where $\hat{s_k}$ denotes the k-th token in the generated summary, and $\hat{s_1}, \hat{s_2}, ..., \hat{s}_{k-1}$ denote previous tokens.

2.7 Model Training

First, we employ the negative log-likelihood loss to supervise the training of the summary generation model, which is formulated as $\mathcal{L}_1 = -\sum_{t=1}^{|\mathcal{S}|} \log(p(s_t|T, V, s_{<t}))$. Second, to train the essential patch selector, we assign a label O to each patch based on oracle creation; set O = 1 if the patch belongs to the oracle, otherwise 0. We then train the model with a binary cross-entropy loss function: $\mathcal{L}_2 = -\frac{1}{n}\sum_{i=1}^{n}[O_i \log(r_i) + (1 - O_i)\log(1 - r_i)]$, where O_i is the label for the i-th patch, and r_i is the relevance score obtained from the MLP. Finally, we expect the model to understand the summary and grasp the vision-summary correlation. To this end, we introduce a task in which the model generates the corresponding summary S directly from visual information, without access to the source text T. This enables our model to develop a preliminary understanding of the summary and grasp the overall context, formulated as $\mathcal{L}_3 = -\sum_{t=1}^{|\mathcal{S}|} \log(p(s_t|\dot{V}, s_{<t}))$.

Overall Loss. The final loss is the combination of the above three losses: $\mathcal{L} = \mathcal{L}_1 + \alpha \cdot \mathcal{L}_2 + \beta \cdot \mathcal{L}_3$, where α and β are hyper-parameters for regulating the balance among the three loss components. More details about the model are in Appendix A.

2.8 Oracle Creation

To facilitate supervised training of the essential patch selector in PRVIN, an oracle label should be assigned to each patch. We introduce a novel method for oracle creation that synergizes image–text similarity with an object detection model. We first select candidate patches with image–text similarity, and then verify them with an object detection model to ensure they contain text-related objects. This approach can mitigate bias by avoiding reliance on a single metric. The details and the illustrations are in Appendix B.

3 Experiments

3.1 Experimental Settings

We evaluated our model on two MMS datasets, MMSS (Li et al., 2018) and MM-Sum-En (Liang et al., 2023). We set Top-k and Top-j to 70% and 60% of all patches, respectively, and set both the balancing factors α and β in the loss function to 1.0, since we found they were optimal on the validation set through tuning with grid search. More details of the datasets, the selection of the ratios and the balancing factors, the discussion of computational efficiency, and the experimental settings are in Appendices C, D, E, and F.

3.2 Experimental Results

We employed six distinct evaluation metrics to rigorously assess the performance of our model;

Model	R-1	R-2	R-L
MMSS			
MAtt	47.28	24.85	44.48
CFSum	47.86	25.64	44.64
*VG-BART**	51.73	29.17	48.91
T-3	53.71	30.96	50.62
VE-ELIN	**54.20**	**31.24**	**51.47**
PRVIN	**54.16**	**31.19**	**51.31**
MM-Sum-En			
mT5	36.99	15.18	29.64
VG-mT5	37.17	14.88	29.41
SOV-MAS	37.26	15.02	29.61
*VG-BART**	37.27	15.91	30.26
VE-ELIN	39.97	18.09	32.47
PRVIN	**40.26**	**18.37**	**32.61**

Model	BLEU	BertScore	MoverScore
MMSS			
CFSum	48.83	86.98	32.36
*VG-BART**	57.63	91.80	45.91
T-3	59.68	91.99	**63.96**
VE-ELIN	**60.16**	**92.22**	49.15
PRVIN	**60.26**	**92.15**	**49.32**
MM-Sum-En			
*VG-BART**	40.88	90.73	27.47
VE-ELIN	45.44	96.61	30.85
PRVIN	**45.90**	**97.04**	**31.18**

Table 1: Experimental results on the test set of the MMSS and MM-Sum-En datasets. Scores with an asterisk (∗) are reproduced scores, otherwise reported scores from the original paper.[2] Our model is statistically significantly better than the underlined scores at p-value < 0.01 with paired bootstrap resampling (Koehn, 2004).

ROUGE-1, 2, L (Lin, 2004), BLEU (Papineni et al., 2002), MOVERScore (Zhao et al., 2019), and BERTScore (Zhang et al., 2019). The details are described in Appendix G. We compared our model with several SOTA models: **CFSum** (Xiao et al., 2023), **VG-BART** (Yu et al., 2021), **T-3** (Yuan et al., 2024), **VE-ELIN** (Yan et al., 2024), **SOV-MAS** (Liang et al., 2023), and several classical methods. The details of these models are described in Appendix H.

Table 1 presents the overall results of our principal metrics on the two datasets. For the MMSS dataset, PRVIN achieved competitive performance to the state-of-the-art VE-ELIN model across all metrics. In particular, it achieved a higher MoverScore (49.32) than VE-ELIN (49.15), indicating better text-generation diversity. Notably, PRVIN outperformed the strong baselines including CFSum and T-3, indicating its effectiveness in leveraging multi-modal information. For the MM-Sum-En dataset, PRVIN outperformed all baselines. This performance advantage indicates strong model robustness across different data types and reveals a distinctive capability in processing lengthy textual content, which we attribute to our architecture's effective integration of multi-modal features. To more comprehensively evaluate our model, we performed comparisons with state-of-the art multimodal large language models, demonstrated the superiority of our model(Appendix I). We also show example outputs from our model in Appendix J.

[2]We attempted to reproduce the strongest baseline but could not due to unavailable code; some models reported only ROUGE. All our results are averaged over 3 trials.

method	R-1	R-2	R-L
MMSS			
Full model	**54.16**	**31.19**	**51.31**
w/o *refiner*	51.74	29.10	49.76
w/o *selector&refiner*	51.03	28.33	48.75
w/o *vision-sum loss*	53.77	31.09	50.93
MM-Sum-En			
Full model	**40.26**	**18.37**	**32.61**
w/o *refiner*	39.05	17.87	31.87
w/o *selector&refiner*	38.12	17.21	31.20
w/o *vision-sum loss*	40.04	18.12	32.33

Table 2: Ablation study results on two datasets.

3.3 Ablation Study

We conducted an ablation study to verify the importance of each module in our model. Specifically, we evaluated PRVIN's performance by removing each component. The results are shown in Table 2. Removing the patch refiner degrades the performance, demonstrating that refining both patch-level visual features and inter-patch correlations enhances PRVIN's effectiveness. Removing the selector and refiner causes the performance to markedly decline, highlighting their essential contribution to text-relevant patch selection and refinement. Furthermore, removing the vision-summary loss harms model performance, highlighting the crucial benefit of direct image-summary alignment. More analysis is found in Appendix K. We also evaluated the impact of removing text and image encoders on the model in Appendix L.

4 Conclusion

We proposed a Patch-Refined Visual Information Network (PRVIN) for solving the MMS problems at the image patch level. Our model consists of an essential patch selector that identifies vision-text-summary correlations and prioritizes semantically critical image patches, and a patch refiner that processes the selected patch sequences to further distill and optimize visual information. We additionally utilize a vision-to-summary auxiliary task that explicitly models vision-summary interdependence. Extensive experiments conducted on two benchmark datasets demonstrated the superiority and effectiveness of our approach.

Limitations

This study is subject to two primary limitations. First, the generalizability of our approach requires further validation due to dataset constraints. The current validation has been restricted to two benchmark datasets, potentially limiting the model's adaptability to other non-english languages and other modalities.

Second, the model performance exhibits significant dependency on the oracle creation method. While our oracle creation framework demonstrates theoretical advancements over conventional methods, its empirical optimality remains unverified due to the absence of systematic evaluation metrics. This methodological uncertainty may impact the model's robustness, particularly when applied to complex real-world scenarios, where oracle quality could substantially affect downstream task performance. We leave these as our future work.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. *arXiv preprint arXiv:1702.01287*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Anubhav Jangra, Sriparna Saha, Adam Jatowt, and Mohammad Hasanuzzaman. 2020. Multi-modal summary generation using multi-objective optimization. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1745–1748.

Nidhal Jegham, Chan Young Koh, Marwan Abdelatti, and Abdeltawab Hendawi. 2024. Evaluating the evolution of yolo (you only look once) models: A comprehensive benchmark study of yolo11 and its predecessors. *arXiv preprint arXiv:2411.00201*.

Rahima Khanam and Muhammad Hussain. 2024. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*.

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

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

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 the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, et al. 2024. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision*, 16(1-2):1–214.

Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, Chengqing Zong, et al. 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.

Yunlong Liang, Fandong Meng, Jinan Xu, Jiaan Wang, Yufeng Chen, and Jie Zhou. 2023. Summary-oriented vision modeling for multimodal abstractive summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2934–2951, Toronto, Canada. Association for Computational Linguistics.

Jindřich Libovickỳ and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. *arXiv preprint arXiv:1704.06567*.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

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

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Shruti Palaskar, Jindrich Libovickỳ, Spandana Gella, and Florian Metze. 2019. Multimodal abstractive summarization for how2 videos. *arXiv preprint arXiv:1906.07901*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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.

AM Rush. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347*.

Min Xiao, Junnan Zhu, Haitao Lin, Yu Zhou, and Chengqing Zong. 2023. Cfsum: A coarse-to-fine contribution network for multimodal summarization. *arXiv preprint arXiv:2307.02716*.

L Xue. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Haolong Yan, Binghao Tang, Boda Lin, Gang Zhao, and Si Li. 2024. Visual enhanced entity-level interaction network for multimodal summarization. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3248–3260.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021. Vision guided generative pre-trained language models for multimodal abstractive summarization. *arXiv preprint arXiv:2109.02401*.

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

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

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *arXiv preprint arXiv:1909.02622*.

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

Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. 2017. Selective encoding for abstractive sentence summarization. *arXiv preprint arXiv:1704.07073*.

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

A Model Details

PRVIN consists of the following components: a 6-layer transformer for the text encoder, a 12-layer transformer for the vision encoder, a 3-layer transformer for the cross-modal encoder, a 3-layer perceptron for the essential patch selector, a 2-layer transformer for the patch refiner, and a 6-layer transformer for the multi-modal decoder. Specifically, we initialized the text encoder and multi-modal decoder using the encoder and decoder of BART-base (Lewis et al., 2020), respectively, with a text feature dimension of 768. The patch refiner was initialized using the last two layers of the BART-base decoder and the cross-modal encoder network was initialized using the last three layers of the BART-base encoder. For the vision encoder, we used the vision encoder from the "ViT-B/32" version of CLIP (Radford et al., 2021), with a patch size of 32×32 and an output visual feature dimension of 768. The essential patch selector was placed between the 8-th and 9-th transformer layers in the vision decoder (V=8).

B Oracle Creation

Our novel oracle creation approach begins with an initial screening based on image–text similarity, followed by the selection of oracle patches using an auxiliary object detection model. Specifically, we first employ CLIPScore (Hessel et al. (2021)) to compute the similarity between each patch and both the input text and the reference summary:

$$Sim_{p_i->T} = CLIP(p_i, T), \quad (1)$$

$$Sim_{p_i->S} = CLIP(p_i, S). \quad (2)$$

The resulting similarity values are then aggregated into a comprehensive similarity sequence, from which patch indices are extracted according to a predefined ratio (20%) to form the oracle, and selected patches are subsequently assigned a label of 1.

To enhance the precision of oracle selection, we introduce a three-stage object detection-assisted refinement framework. The pipeline operates as follows: **Object Localization**: We employ YOLOv11m (Khanam and Hussain, 2024) as our pre-trained object detection backbone due to its optimal trade-off between model compactness, computational efficiency, and detection accuracy (Jegham et al., 2024). This network generates object category prediction along with the corresponding bounding box for each image.

Figure 2: Original images and their corresponding oracle patches

Semantic-relevance Filtering: For each detected object, we verify its semantic alignment with the target summary through synonym matching. Specifically, we maintain a synonym set derived from WordNet[3] for each object category. An object is considered relevant if any lexical item from its synonym set appears in the target summary. Images without relevant objects retain their original CLIPScore evaluation.

Oracle-region Optimization: When objects are deemed relevant, we calculate the proportion of image area occupied by their bounding boxes. If the cumulative area exceeds 20% of the total image space, we strategically select the most representative 20% of patches (prioritizing central regions of large objects). For sub-20% coverage cases, we preserve all object-associated patches. The remain-

[3] www.nltk.org

ing oracle patches are then supplemented from the highest-ranked CLIPScore regions to maintain consistent selection quantities.

This hybrid approach synergistically combines semantic understanding from object detection with cross-modal alignment from CLIPScore, ensuring both semantic relevance and visual-textual correspondence in oracle selection. As shown in the oracle examples in Figure 2, our method successfully identifies patches closely aligned with the summary content.

C Datasets

We conducted experiments on the representative MMSS dataset (Li et al., 2018), which contains 62,000/2,000/2,000 samples for the training/validation/test set, respectively. Each sample in the dataset is a triplet $<sentence, image, summary>$. We also evaluated our method on the English part of the multilingual multi-modal abstractive summarization (MM-Sum-En) dataset (Liang et al., 2023), which contains 326,725 samples and 867,817 images in total, all sourced from BBC News. Using a 93%/3.5%/3.5% split, the training, validation, and test sets contain 303,828, 11,437, and 11,460 samples, respectively. Each sample consists of a triplet: $<news\ article, associated\ images, summary>$.

D The Impact of Selection Ratios and Balancing Factors

To investigate the impact of selection ratios in the essential patch selector and balancing factors in the loss function on PRVIN's performance, we tried different combinations of k, j, α, and β for evaluation.

For the balancing factors, we searched the optimal α and β for each combination of k and j among the values of $\{0.25, 0.5, 0.75, 1.0, 1.25\}$. Eventually, we found that our model is not sensitive to α and β because the differences were less than 0.05, and so we set both α and β to 1.0.

For the selection ratios, as shown in Figure 3, PRVIN's performance gradually improved with increasing k values, but it stagnated significantly when reaching approximately 70% of image information utilization. Furthermore, PRVIN's performance also gradually improves with increasing j values and then tends to stabilize. When the j value is too high, for example, exceeding 0.8, the model performance decreases. We conjecture that

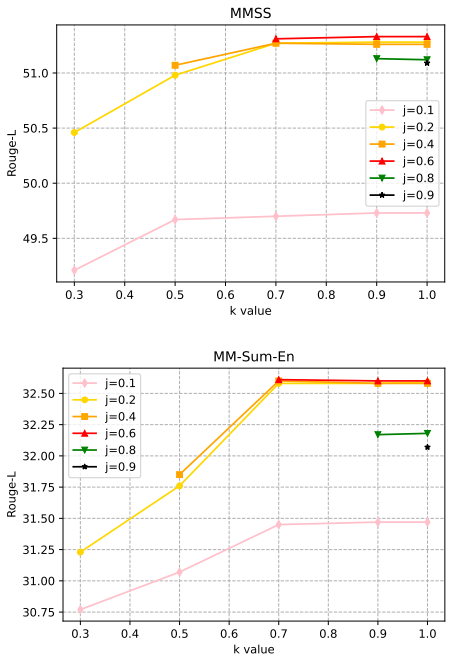

Figure 3: The impact of different ratios for k and j on model performance on the validation set of the MMSS and MM-Sum-En datasets.

redundant patches not only prove unproductive but may potentially compromise summarization quality through information dilution. We obtained similar results on the MM-Sum-EN validation set.

E Efficiency of Patch Selector and Patch Refiner

To validate the efficiency of the proposed two-stage framework comprising a patch selector and a patch refiner, in Figure 4, we report inference latency as a measure of computational speed under different selection ratios: k for selector and j for the patch refiner, thereby quantifying the efficiency gains achieved by our two-stage framework. The results indicate a consistent reduction in latency as the proportion of selected patches decreases, evidencing the patch selector's ability to lower computational demand. Theoretically, since the vision encoder receives substantially fewer visual tokens and PRVIN's backbone is composed of transformer blocks whose complexity grows with sequence length, reducing the token count yields considerable savings in FLOPs. Combining the evidence from Figures 3 and 4, we observe that selective pruning of redundant visual tokens reduces computational cost while maintaining comparable summarization performance.

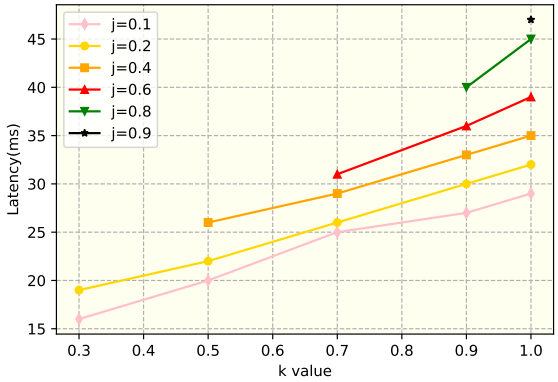

Figure 4: Comparison of the efficiency for different selection ratios on the MM-Sum-En dataset

F Training Details

For the MMSS dataset, we set the batch size to 16, the dropout to 0.1, the maximum training epochs to 50, and the beam size to 10. The model was optimized using Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.998$, and the learning rate was set to 5e-6. The maximum input length was 64 and the maximum output length was 32. For the MM-Sum-EN dataset, the parameters remained identical to those in the MMSS, except that the maximum input length was 1024 because of the limitation of BARTbase, the maximum output length was 256, the batch size was 8, and the maximum training epochs was 20 (Liang et al., 2023). All models were trained and tested on a two A100-80GB GPU.

G Evaluation Metrics

Following previous studies, we presented our experimental results in terms of 6 automatic metrics: ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004) as principal metrics, and BLEU (Papineni et al., 2002), MOVERScore (Zhao et al., 2019), and BERTScore (Zhang et al., 2019) as supplementary metrics to ensure a comprehensive assessment. BLEU computes the n-gram precision between a candidate and references, with a brevity penalty to discourage overly short outputs. BERTScore leverages contextual embeddings from BERT (Devlin, 2018) by calculating the cosine similarity between candidate and reference tokens. MoverScore measures semantic distance using the Word Mover's Distance (Kusner et al., 2015) between the distributions of word embeddings in the candidate and reference texts, enabling a robust evaluation of meaning.

H Compared Methods

To evaluate the effectiveness of our method, we compared it against some classic and strong baselines. For the MMSS dataset: **Lead** generates summaries by directly extracting the first eight words of the source. **Compress** employs an integer linear programming framework to achieve sentence compression, leveraging the syntactic structure as its foundational basis. **ABS** (Rush, 2015) utilizes an attention-based CNN encoder paired with a neural language model decoder to produce abstractive summaries. **SEASS** (Zhou et al., 2017) introduces a selective encoding mechanism that dynamically filters and prioritizes salient textual features during the summarization process. **Multi-source** (Libovický and Helcl, 2017) introduces flat and hierarchical attention strategies to integrate multiple source modalities. **Doubly-Attentive** (Calixto et al., 2017) utilizes a doubly-attentive mechanism to incorporate visual features. Table 3 shows the comparison of classic methods and our model. **MAtt** (Li et al., 2018) proposes a modality-based attention mechanism and an image filter to enhance the relation between modalities. **VG-BART** (Yu et al., 2021) utilizes PLMs as the backbone and injects visual features into the encoder layer through dot production. **CFSum** (Xiao et al., 2023) proposes a contribution network to calculate image contributions and guide the attention of both textual and visual modalities. **T-3** (Yuan et al., 2024) resorts Information Bottleneck (IB) to alleviate over-preservation and over-compression of visual information. **VE-ELIN** (Yan et al., 2024) considers entity-level granularity to address the problem of under-utilization of multi-modal inputs and is a strong baseline.

For the MM-Sum-En dataset, **mT5** (Xue, 2020) uses a multilingual variant of T5, that was pretrained on a new Common Crawl-based dataset covering 101 languages, and is the text-only baseline. **VG-mT5** (Liang et al., 2023) implements an attention based text-vision fusion method to inject visual features into the mT5 model. **SOV-MAS** (Liang et al., 2023) proposes two summary-oriented auxiliary tasks to enhance the MAS model based on the pre-trained language models.

I Comparison with MLLM-based Methods

Building on the remarkable achievements of Multimodal Large Language Models (MLLMs) across

Model	R-1	R-2	R-L
	MMSS		
Lead†	33.64	13.40	31.84
Compress†	31.56	11.02	28.87
ABS†	35.59	18.21	31.89
SEASS†	44.86	23.03	41.92
Multi-Source†	39.67	19.11	38.03
Doubly-Attentive†	41.11	21.75	39.92
PRVIN	**54.16**	**31.19**	**51.31**

Table 3: Comparison with some classic baselines, The results marked with "†" are reported in (Li et al., 2018).

Model	**R-1**	**R-2**	**R-L**
UniG †	46.22	24.28	43.47
BLIP-MMSS †	48.43	26.76	45.87
zero-shot			
LLaVA-v1.6-Mistral-7B	15.72	4.02	14.01
Qwen2.5-VL-7B	18.81	6.10	14.95
PRVIN	**54.16**	**31.19**	**51.31**

Table 4: Comparision with MLLM-based methods on the test set of MMSS. The results marked with † are reported in (Yuan et al., 2024).

diverse multi-modal tasks (Yin et al., 2023; Li et al., 2024), we further expanded our comparative analysis to include state-of-the-art MLLM-based approaches. To establish a comprehensive benchmark, we implemented a zero-shot evaluation protocol using two leading MLLMs, LLaVA-v1.6-Mistral-7B (Liu et al., 2024) and Qwen-2.5-VL-7B (Bai et al., 2025). Specifically, we prompted LLaVA-v1.6-Mistral-7B and Qwen-2.5-VL-7B with the following prompt to generate a multi-modal summary for the MMSS dataset: *Combine the following text with the image content, summarize coherently including content from both the text and the image, and compress to one sentence such that it captures the most salient information from both modalities.*

Following Yuan et al. (2024), we also conducted comparison against vision-language pre-trained (VLP) models; UniG (Xiao et al., 2023) adopts an identical architectural framework to CFsum, with the primary distinction residing in its encoder component, which utilizes the UNITER model (Chen et al., 2020). The BLIP-MMSS framework is constructed through fine-tuning the BLIP (Li et al., 2022) model on the MMSS dataset. As shown in Table 4, our model clearly far outperforms the zero-shot capabilities of the current top-tier multi-modal large models, indicating that further sophisticated designs and approaches are needed for these models to be effectively applied to multi-modal summarization tasks. Moreover, our model also outperforms vision pre-trained model-based methods, such as UniG and BLIP-MMSS, which demonstrates its more efficient utilization of visual information.

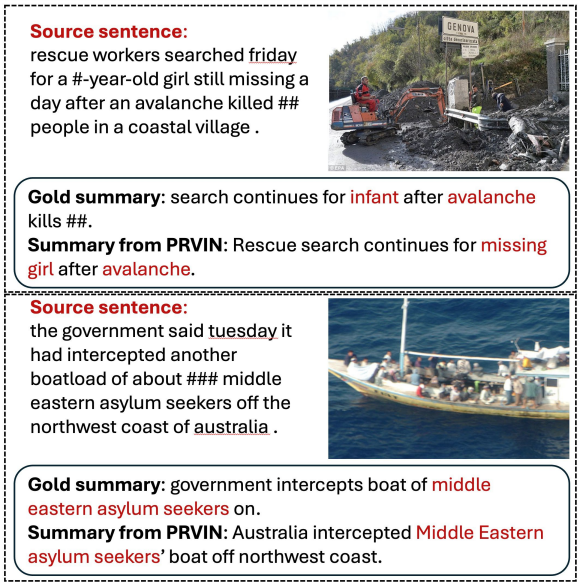

Figure 5: Case study for our PRVIN

method	**R-1**	**R-2**	**R-L**
Full model	54.11	31.17	51.27
w/o text encoder	49.12	28.07	47.91
w/o vision encoder	51.51	28.63	48.72

Table 5: The effectiveness of text and vision encoders on the test set of MMSS

J Case Study

Figure 5 shows examples of multi-modal summary generation outputs using our model.

K Impact of Vision-to-Summary Loss

We conducted a case study on the vision-to-summary (vis2sum) loss. As shown in Figure 6, the vis2sum loss improved the model's visual understanding (e.g., correctly recognizing "overpass"). Nevertheless, ablation study results show only marginal ROUGE gains (+0.4). We hypothesize that, since the proposed patch selector and patch refiner already extract and integrate essential visual content effectively, the task of directly generating summaries from images provides limited additional benefit to final performance.

Source: this ##th century town in chile 's bucolic central valley was especially hard-hit by the powerful #.# magnitude earthquake that struck the country on saturday .
Summary: quake slams ##th century chilean town, overpass falls.

Full model:
Earthquake devastates Chillán, Chile; overpass collapses.
w/o v2s loss:
Quake devastates Chillán, Chile on Saturday.

Figure 6: Case study for our vision-to-summary loss

L The Effectiveness of the Text and Vision Encoders

To validate whether PRVIN genuinely leverages multi-modal information rather than relying on a single modality, we conducted ablation studies by masking specific components for the cross-modal representation. Specifically, as shown in Table 5, "w/o text encoder" indicates masking the textual portion of the cross-modal representation, while "w/o vision encoder" refers to masking the visual counterpart. Our experimental results reveal two key findings. First, when removing the text encoder, PRVIN still generates summaries of reasonable quality, which we attribute to the effectiveness of our vision-sum loss in maintaining visual semantic preservation. Second, removing either text or vision encoder causes a significant performance degradation, demonstrating that PRVIN effectively exploits information from both modalities. This substantial performance difference between single-modality and dual-modality ablation confirms that PRVIN successfully captures cross-modal correlations, enabling synergistic integration of textual and visual information to enhance summarization quality. The empirical evidence suggests that, while each modality contributes independently, their coordinated interaction through our concatenated representation learning yields optimal performance.