Language Constrained Multimodal Hyper Adapter For Many-to-Many Multimodal Summarization

Nayu Liu[1], **Fanglong Yao**[3], **Haoran Luo**[4], **Yong Yang**[1], **Chen Tang**[5], **Bo Lv**[2]*
[1]Tianjin Key Laboratory Autonomous Intelligence Technology and Systems, School of Computer Science and Technology, Tiangong University, [2]University of Chinese Academy of Sciences
[3]Key Laboratory of Target Cognition and Application Technology, AIRCAS
[4]Graduate School of Creative Science and Engineering, Waseda University
[5]Institute for Advanced Algorithms Research, Shanghai
nayuliu@tiangong.edu.cn, lvbo19@mails.ucas.ac.cn

Abstract

Multimodal summarization (MS) combines text and visuals to generate summaries. Recently, many-to-many multimodal summarization (M3S) garnered interest as it enables a unified model for multilingual and cross-lingual MS. Existing methods have made progress by facilitating the transfer of common multimodal summarization knowledge. While, prior M3S models that fully share parameters neglect the language-specific knowledge learning, where potential interference between languages may limit the flexible adaptation of MS modes across different language combinations and hinder further collaborative improvements in joint M3S training. Based on this observation, we propose Language Constrained Multimodal Hyper Adapter (LCMHA) for M3S. LCMHA integrates language-specific multimodal adapters into multilingual pre-trained backbones via a language constrained hypernetwork, enabling relaxed parameter sharing that enhances language-specific learning while preserving shared MS knowledge learning. In addition, a language-regularized hypernetwork is designed to balance intra- and inter-language learning, generating language-specific adaptation weights and enhancing the retention of distinct language features through the regularization of generated parameters. Experimental results on the M3Sum benchmark show LCMHA's effectiveness and scalability across multiple multilingual pre-trained backbones.

1 Introduction

Multimodal summarization (MS) (Zhang et al., 2024; Zhu et al., 2020; Jangra et al., 2023; Mahon and Lapata, 2024; Krubiński and Pecina, 2023; Xiao et al., 2023; Zhuang et al., 2024) aims to generate summaries based on multimodal documents such as text and visuals, helping people quickly locate key information from the vast multimedia content on the internet. Recently, many-to-many multimodal summarization (M3S) (Liang et al., 2023a; Verma et al., 2023) that incorporates multilingual MS and cross-lingual MS has gained the interests. As shown in Figure 1 (a) and (b), M3S allows a unified model to handle multimodal summarization for different language combinations, which enhances the efficiency of MS model deployment in multiple language scenarios (Ghosh et al., 2024; Nguyen et al., 2023; Liang et al., 2023b).

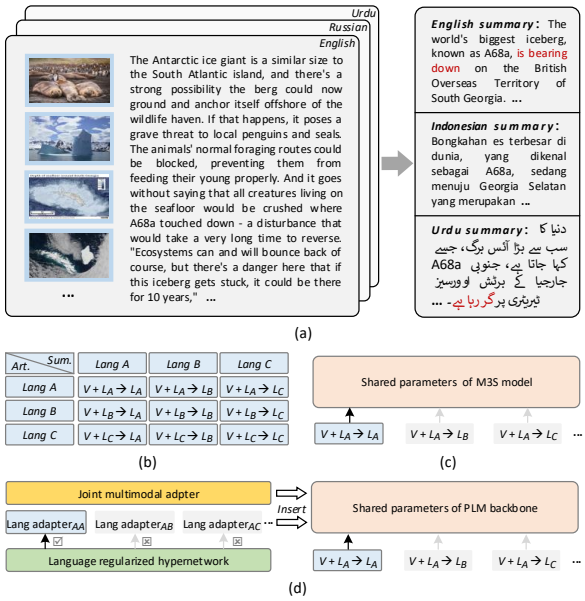


Figure 1: (a) An M3S example. The words marked in red in Urdu (translated as "bearing down") appears at the end of the sentence, following the subject-object-verb pattern, in contrast to the subject-verb-object pattern in English. (b) M3S integrates visual (V) and article-summary across languages (e.g. L_A, L_B, L_C). (c) Previous M3S models that fully shared parameters. (d) LCMHA reconcile common multimodal and language-specific summarization knowledge learning by language-specific multimodal hyper adpaters.

A key challenge in M3S lies in balancing the learning of common multimodal summarization knowledge and language-specific knowledge. Re-

*Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 25285–25298

cent works improve M3S by facilitating the transfer of common multimodal summarization knowledge. For example, Liang et al. (2023b) guided summarization through a rich shared visual domain beyond the impact of different language articles. Shi (2024) utilized one unified decoder with sequential learning to improve cross-lingual alignment. Liu et al. (2022b) and Liang et al. (2023a) introduced cross-lingual distillation strategies to bridge the gap between high- and low-resource language features, enhancing positive knowledge transfer between languages. It is reasonable to expect that leveraging shared visual information and cross-lingual transfer strategies could effectively benefit summarization across different languages. Nevertheless, previous methods that focus on common MS knowledge transfer may neglect the learning of language-specific knowledge, where using fully shared parameters to learn M3S models may cause potential inter-language interference (Wang et al., 2020; Huang et al., 2023), limiting the flexible adaptation of MS modes across different language combinations and hindering further collaborative improvements. We aim to maintain the learning of common visual-language knowledge while preserving the independent knowledge learning of different language combinations.

Motivated by the analysis of these reasons, we propose Language Constrained Multimodal Hyper Adapter (LCMHA) for M3S, integrating language-specific multimodal adapters via a hypernetwork into multilingual pre-trained backbones, to address the flexible adaptation of diverse knowledge in the joint MS training in different language combinations. As shown in Figure 1 (d), LCMHA relaxes language parameter sharing to improve independent language knowledge learning via language-specific adapter sets, while maintaining the interaction of language features from various adapters and common visual features via a shared multimodal adapter module. In addition, a hypernetwork constrained by a language regularization term is designed to generate language-specific adaption weights based on source and target languages, and enhance the retention of characteristics for each language combination of M3S by the regularization of generated parameters.

Experimental results on the M3Sum benchmark (Liang et al., 2023a) show that the proposed LCMHA method outperforms related approaches, and we also replace different multilingual pretrained language model (PLM) backbones including mT5 (Xue et al., 2021), mBART (Tang et al., 2021) and LLaMA (Touvron et al., 2023) to verify the scalability of our approach. Moreover, the proposed method demonstrates the effectiveness of using language-specific modules and a shared multimodal module to improve M3S. Our contributions are as follows:

- We propose LCMHA to facilitate flexible adaptation in multilingual/cross-lingual M3S by integrating language-specific multimodal hyper adapters, preserving both shared visual-language and independent language combination knowledge learning.
- We introduce a language-regularized hypernetwork to balance intra- and inter-language learning, generating language-specific adapter weights and preserving distinct language features.
- The proposed method surpasses previous state-of-the-art methods on different multilingual pretrained backbones, showcasing the benefits of using language-specific and shared multimodal modules to improve M3S[1].

2 Related Work

2.1 Multimodal Summarization

Multimodal summarization (MS) seeks to compress multimodal data (e.g. text, images, videos, etc.) to generate summaries (Patil et al., 2024; Hirao et al., 2024; Chen and Zhuge, 2018; Li et al., 2017, 2018; Qiu et al., 2022; Liu et al., 2022a; He et al., 2023; Zhang et al., 2022b,a; Im et al., 2021; Shang et al., 2021; Jiang et al., 2023; Xiao et al., 2024; Zhu et al., 2021), spanning various scenarios including news (Fu et al., 2021), social media (Overbay et al., 2023), meetings (Li et al., 2019), e-commerce (Li et al., 2020), and videos (Khullar and Arora, 2020; Tang et al., 2023). Existing works mainly focus on efficiently utilizing multi-source information and well-designed multimodal interaction schemes (Palaskar et al., 2019; Liu et al., 2020; Yu et al., 2021; Lin et al., 2023; Yan et al., 2024; Fan et al., 2025; Xiao et al., 2025).

MS in multilingual and cross-lingual scenarios (Nguyen et al., 2023; Hasan et al., 2021; Liu et al., 2022b; Wang et al., 2022; Ghosh et al., 2024; Shi,

[1]The code is released at https://github.com/lvbotenbest/HAMMS

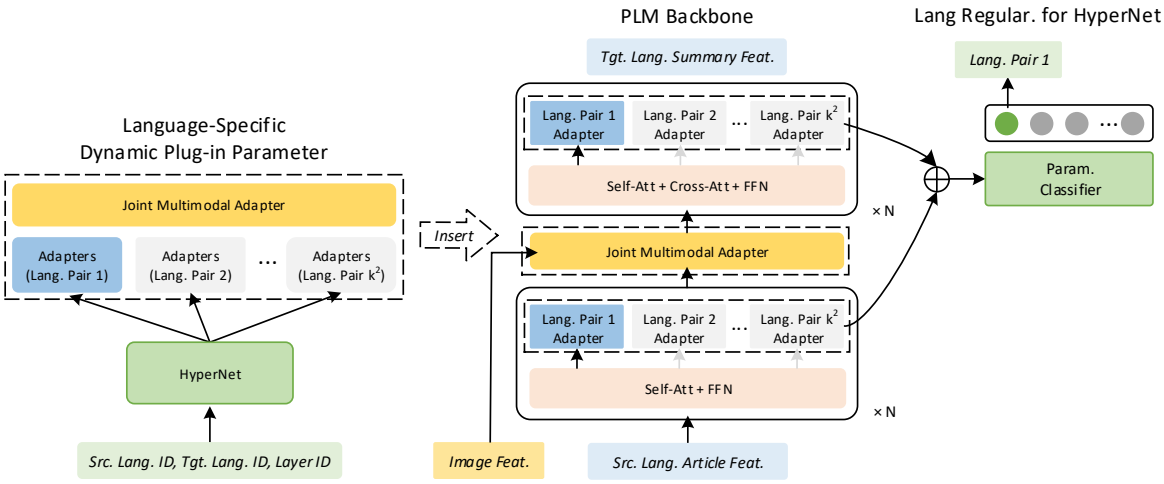


Figure 2: The architecture of LCMHA. A language constrained hypernetwork generates language-specific adapters based on source article and target summary languages and layer IDs, and together with the joint multimodal adapter, are integrated into multilingual PLM backbones. The accumulated adaptation parameters perform language pair classification as the language regularization of hypernetwork, co-optimizing the model with the summarization loss in a multi-task learning paradigm. The figure illustrates an encoder-decoder architecture as an example; for decoder-only architectures, the multimodal adapter is placed within the intermediate Transformer blocks.

2024; Liu et al., 2024) has gained increasing attention, as most prior works primarily focus on monolingual settings. Recent studies explore visual guidance and cross-lingual knowledge transfer. Liang et al. (2023b) proposed two image-guided auxiliary training tasks that leverage shared image information across multiple languages to enhance mid- high-, low-, and zero-resource language summarizations. Nguyen et al. (2023) introduced visual layout information to capture long-range dependencies in lengthy text inputs. Liu et al. (2022b) and (Liu et al., 2024) introduced cross-lingual knowledge distillation strategies (Lv et al., 2024) to transfer knowledge from the monolingual MS model trained with rich resources to assist cross-lingual MS. Liang et al. (2023a) first introduced the M3S task, incorporating cross-lingual dual distillation and image-summary contrastive learning to improve summarization. Verma et al. (2023) released an M3S dataset covering 8 high-resource and 12 low-resource languages, reporting popular baseline results such as mT5 (Xue et al., 2021) and MSMO (Zhu et al., 2018). Despite the progress in prior methods, the neglected language-specific knowledge learning may limit their flexibility in adapting to different summarization patterns across various language combinations. To address this, we propose LCMHA, which first explores the use of hypernetworks and adapters in MS to reconcile common visual-language knowledge learning and independent article-summary adaptation across different language combinations.

2.2 Adapter & Hypernetwork

Adapters (Rebuffi et al., 2017) are lightweight modules integrated into the backbone model, allowing it to be fine-tuned for specific tasks. Adapter technology has demonstrated its superiority in various domains (Bapna and Firat, 2019; Feng et al., 2024; Wang et al., 2021; Sung et al., 2022; Hu et al., 2023). Our work extends adapter-based approaches to improve knowledge transfer in M3S.

Hypernetworks (Ha et al., 2022) primarily function to generate weights or adapters for another network, aiming to adaptively adjust the parameters of the main model, thereby providing greater flexibility and generalization capability (Mahabadi et al., 2021; Tay et al., 2020). They have been applied in continual learning (Von Oswald et al., 2020), multi-task learning (Ivison et al., 2023; Zhao et al., 2024), and cross-lingual transfer (Ansell et al., 2021; Baziotis et al., 2022; Üstün et al., 2022).

3 Method

3.1 Overview

In a given language collection, specifying a source language multimodal document as input, the goal of the M3S system is to generate abstractive summaries in the specified target language. Formally, let the language set be represented as L_1 to L_k, and the IDs of the source and target languages

be L_{src} and L_{tgt} respectively, where $L_{src}, L_{tgt} \in \{L_1, ..., L_k\}$. Given a source language article input $X = \{x_1, ..., x_{|X|}\}$ and the corresponding image sets $V = \{v_{11}, ..., v_{1n}, ..., v_{m1}, ..., v_{mn}\}$, where x_i denotes the i-th token in the article sequence of length $|X|$, and v_{ij} denotes the j-th object in the i-th image. The target language summary is represented as $Y = \{y_1, ..., y_{|Y|}\}$. The M3S task is formally defined as:

$$\arg\max_{\theta} \prod_{t=1}^{|Y|} P(y_t|X, V, L_{src}, L_{tgt}, y_{<t}; \theta) \quad (1)$$

where θ is trainable parameters. In the following section, we first introduce the internal structure of LCMHA, including 1) Language-specific multimodal adapters, which are integrated into the pre-trained backbone model to learn language-specific knowledge and common multimodal summarization knowledge; and 2) Language regularized hypernetwork, which generates adaption weights to balance intra- and inter-language learning. We then describe integrating LCMHA into the pre-trained model for training and inference.

3.2 Language-specific Multimodal Adapters

As shown in Figure 2, each source language article-target language summary combination is assigned independent adaptation parameters to relax multi-language weight sharing to preserve each language's unique characteristics learning. These follow a series adapter (Hu et al., 2023) setting, which consists of an up-projection linear layer and a down-projection linear layer with a ReLU activation function, uniformly connected to each layer of the backbone, denoted as:

$$\text{AdaperL}(X) = W_{up}\text{ReLU}(W_{down}X + b_{down}) + b_{up} \quad (2)$$

where $W_{up} \in \mathbb{R}^{d_b \times d}$, $b_{up} \in \mathbb{R}^d$, $W_{down} \in \mathbb{R}^{d \times d_b}$, and $b_{down} \in \mathbb{R}^{d_b}$ denote the up-projection and down-projection layer parameters respectively, and X denotes the input text feature sequence.

For visual-language fusion, we apply a joint multi-modal adapter module to absorb common visual knowledge beyond different languages. Instead of conventional linear projection adapters that are not proficient at handling multimodal interactions, we use a forget gate fusion module to build the multimodal adapter, integrating features from different language pair adapter branches and visual features, denoted as:

$$\text{AdapterM}(X, V) = W\text{Concat}(M \otimes G, X) + b \quad (3)$$

$$M = \text{Att}_{X \to V}(X, V) = \text{Softmax}(XW_m V^T)V \quad (4)$$

$$G = \text{Sigmoid}(W_g\text{Concat}(X, V) + b_g) \quad (5)$$

where W, b, W_m, W_g, b_g are learnable adaption parameters, $\text{Concat}(\cdot)$ denotes concatenation along the feature dimension, X denotes the features output by the language adapters, V denotes the visual features, M denotes the cross-modal visual context features attended by the text, and G denotes the forget gate that filters cross-modal context noise.

For example, in the English-Urdu MS scenario, the input information is processed through the backbone model and the English-Urdu language adapter branch, with all language branches being aggregated by the joint multimodal adapter. The specific weights for different language combinations are not derived from the model's initialization parameters but are generated by the hypernetwork described in the next section, which helps to avoid the high parameter costs associated with adapters for multiple language combinations.

3.3 Language Regularized Hypernetwork

To balance the independent and common knowledge learning across different language combinations, a language regularized hypernetwork (LRH) is introduced to generate language-specific adaptation weights. LRH comprises a shared hypernetwork module and a parameter classifier. Concretely, given the source language, target language, and layer IDs, they are transformed into learnable embeddings, denoted as $e_{L_{src}}, e_{L_{tgt}}, e_{layer} \in \mathbb{R}^{d_e}$. The hypernetwork module aims to generate adapter parameters corresponding to the language combination and the specific layer number based on these embedding inputs. The hypernetwork is composed of stacked feed-forward layers with layer normalization and residual connections, represented as:

$$h_0^{(j)} = \text{Linear}(\text{Concat}(e_{L_{src}}, e_{L_{tgt}}, e_{layer})) \quad (6)$$

$$h_i^{(j)} = \text{Linear}(\text{ReLU}(\text{Linear}(\text{LN}(h_{i-1})))) + h_{i-1} \quad (7)$$

where $\text{Linear}(\cdot)$ denotes a linear layer projection, $\text{LN}(\cdot)$ denotes a layer normalization, and $h_i^{(j)}$ represents the hidden state features output by the i-th layer of the hypernetwork with depth l for the j-th layer adapter of the backbone model. The last layer features $h^{(j)}$ are projected through four linear layers, denoted as:

$$W_{up}^{(j)} = \text{Linear}_{W_{up}}(h^{(j)}) \quad (8)$$

$$b_{up}^{(j)} = \text{Linear}_{b_{up}}(h^{(j)}) \quad (9)$$

$$W_{down}^{(j)} = \text{Linear}_{W_{down}}(h^{(j)}) \quad (10)$$

$$b_{down}^{(j)} = \text{Linear}_{b_{down}}(h^{(j)}) \quad (11)$$

The output features $W_{up}^{(j)} \in \mathbb{R}^{1\times(d_b\times d)}$, $b_{up}^{(j)} \in \mathbb{R}^{1\times d}$, $W_{down}^{(j)} \in \mathbb{R}^{1\times(d\times d_b)}$, $b_{down}^{(j)} \in \mathbb{R}^{1\times d_b}$ are reshaped into the parameters of the up-projection and down-projection linear layers of the adapter in the j-th layer of the backbone, and then inserted into the backbone.

A language regularization term in LRH is designed to enhance the preservation of language-specific feature learning of adaptation weights. Concretely, given the up-projection and the down-projection weights of adapters inserted into the backbone, a language classifier is introduced to classify these generated parameters, represented as:

$$Y^{up} = \text{Softmax}(\text{Linear}_u(\frac{1}{l}\sum_{j=1}^{l} W_{up}^{(j)})) \quad (12)$$

$$Y^{down} = \text{Softmax}(\text{Linear}_d(\frac{1}{l}\sum_{j=1}^{l} W_{down}^{(j)})) \quad (13)$$

where $Y^{up} = \{y_1^u, ..., y_k^u\}$ and $Y^{down} = \{y_1^d, ..., y_k^d\}$ represent the probability distributions of the language pair for accumulated up-projection and down-projection layer parameters. This aims to conduct distinct language weights by classifying the generated adaption parameters. Then the classification loss and the summary generation loss are jointly optimized in a multi-task learning paradigm.

3.4 Absorb LCMHA into Backbone

As shown in Figure 2, the language-specific multimodal adapters via LRH are connected to the feed-forward layers of Transformer blocks in a series adapter pattern. In this work, we follow previous works by using multilingual PLMs, mT5 (Xue et al., 2021) and mBART (Tang et al., 2021) as backbones for integrating the LCMHA module. We also integrate LCMHA into LLaMA (Touvron et al., 2023), and considering its decoder-only architecture, we place the multimodal adapter in the intermediate Transformer blocks.

In terms of visual features, we follow previous work (Liang et al., 2023a,b) to use Faster R-CNN (Ren et al., 2015) to extract visual features of the image set. Specifically, for each image, Faster R-CNNs extract a set of potential object bounding boxes (i.e., Regions of Interest, RoIs) and generate feature maps. These RoI feature maps are pooled and mapped to feature vectors via fully connected layers. The extracted features of the image set are represented as $V = \{v_{11}, ..., v_{1n}, ..., v_{m1}, ..., v_{mn}\} \in \mathbb{R}^{(m\times n)\times d_v}$, where m is the number of images in a sample, n is the number of RoIs extracted from each image, and d_v is the feature dimension of each RoI. The RoI bounding box coordinates, image id embeddings, and region id embeddings are added to the visual features to encode the spatial location information of the image set, represented as:

$$v_{ij} \leftarrow v_{ij} + E_{ij}^{box} + E_i^{img} + E_j^{reg} \quad (14)$$

3.5 Training and Inference

During training, the LCMHA model is optimized by minimizing the cross-entropy loss between the generated and the golden summaries, along with the loss from the hypernetwork language regularization term, formulated as:

$$\mathcal{L} = \mathcal{L}_{sum} + \alpha\mathcal{L}_{cls} \quad (15)$$

$$\mathcal{L}_{sum} = -\sum_{t=1}^{T} \log P(y_t|\hat{y}_{<t}, X, V, L_{src}, L_{tgt}) \quad (16)$$

$$\mathcal{L}_{cls} = -(\sum_{i=1}^{K} y_i^u \log\hat{y_i^u} + \sum_{i=1}^{K} y_i^d \log\hat{y_i^d}) \quad (17)$$

where α is a hyperparameter. LRH are trained together with the backbone in our work, rather than focusing on the parameter efficient fine-tuning, where the generation of these language-specific parameters aims to relax knowledge sharing to address potential differences between languages. During inference, the language-specific multimodal hyper adapters via LRH, along with the PLM backbone, are used to generate summaries. The language regularization module of LRH does not participate in inference.

Language	English	Indonesian	Russian	Urdu
English	24,768	10,037	9,076	6,297
Indonesian	9,814	23,176	7,260	6,324
Russian	8,902	7,329	21,036	5,179
Urdu	6,052	5,810	4,700	17,800

Table 1: Statistics of the M3Sum dataset.

4 Experiments

4.1 Dataset

We use the M3Sum benchmark dataset (Liang et al., 2023a) to construct experiments to evaluate the proposed method. The M3Sum dataset is constructed based on CrossSum (Bhattacharjee et al., 2023) and MM-Sum (Liang et al., 2023b). The complete M3Sum covers 44*44 language directions, and the average article length and summary length across all languages are 520 and 84, respectively, with an average of 3.23 images per article-summary pair. Currently, M3S data for 4*4 language directions are publicly available, including English, Indonesian, Russian, and Urdu. The dataset is divided into 80% training set, 10% validation set and 10% test set. Following previous works, we use these 16 language pairs data to evaluate the proposed method. The statistics are shown in Table 1.

4.2 Setup and Metrics

Implementation Details. The proposed LCMHA is integrated with two multilingual pre-trained models, mT5[2] (Xue et al., 2021) and mBART[3] (Tang et al., 2021), as used in previous works (Liang et al., 2023a,b; Bhattacharjee et al., 2023), as well as an LLM backbone LLaMA-3-8B[4] (Touvron et al., 2023). Please refer to the Appendix A for the full details on data processing, training, and inference hyperparameters.

Metrics. Following Liang et al. (2023a); Wang et al. (2023); Liang et al. (2023b), we use ROUGE-1, 2, L evaluation metrics[5] (Lin, 2004). In addition, GPT-4 is employed for a comprehensive assessment of generated summaries across four aspects: consistency, relevance, coverage, and fluency. The detailed template for GPT-4 evaluations is provided in Appendix B.

4.3 Comparison Models

The following baseline models are used for comparison: 1) **VG-mT5** (Liang et al., 2023b), a multimodal summarization baseline that integrates visual information into mT5 via a forget gate fusion module. 2) **D2TV** (Liang et al., 2023a), built on mT5 and mBART backbones respectively, leveraging contrastive learning between vision and summary to guide summarization beyond different language articles, and utilizing parallel language summary data to promote cross-language knowledge transfer via a cross-lingual dual distillation strategy. 3) **D2TV (w/o KD)** and **D2TV (V-KD)**, derived from D2TV, where the former does not introduce additional parallel language summary data for distillation, and the latter adopts a vanilla unidirection distillation, and both of them do not utilize vision-summary contrastive learning. 4) **LLaMA-3-8B** (Touvron et al., 2023), an LLM trained with LoRA (Hu et al., 2022) for comparison with the proposed method.

4.4 Overall Performance

The overall experimental results of different models are shown in Table 2. Overall, the proposed method substantially outperforms all compared models on different PLM backbones. In terms of the average score of ROUGE-1/2/L, LCMHA exceeds the previous SOTA method D2TV by 1/0.49/0.82 points in the English direction, 1.12/0.59/0.96 points in the Indonesian direction, 0.67/0.24/0.58 points in the Russian direction, and 0.77/0.55/0.69 points in the Urdu direction on the mT5 backbone; and by 2.01/1.02/1.74 points in the English direction, 1.90/1/1.63 points in the Indonesian direction, 2.03/0.95/1.7 points in the Russian direction, and 1.74/1/1.41 points in the Urdu direction on the mBART backbone. Moreover, the proposed method does not leverage the parallel bilingual summary data for cross-language distillation to train the model, indicating the advantages of shared multimodal and independent language hyper adapters of LCMHA to improve M3S.

It can also be observed that while LLaMA achieves a notable improvement in English-, Indonesian-, and Russian-related evaluations, its performance in Urdu-related directions is modest, likely due to the lack of extensive pretraining on low-resource languages. Notably, LCMHA with LLaMA as the backbone demonstrates a clear advantage in the Urdu language directions, as the proposed method not only facilitates the transfer of shared knowledge but also preserves Urdu-specific knowledge learning, preventing interference from high-resource languages. For example, LCMHA achieves improvements of 2.83/1.31/2.11 points on English-Urdu, and 4.97/1.48/2.91 points on Urdu-Urdu.

[2] https://huggingface.co/google/mt5-base

[3] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt

[4] https://huggingface.co/meta-llama/Meta-Llama-3-8B

[5] https://github.com/csebuetnlp/xl-sum/tree/master/multilingual_rouge_scoring

Src \ Tgt	Models	English	Indonesian	Russian	Urdu	Avg.
English →	VG-mT5	35.37 / 12.71 / 27.33	28.42 / 9.83 / 23.18	25.42 / 8.85 / 20.42	31.83 / 10.49 / 25.51	30.26 / 10.47 / 24.11
	D^2TV (w/o KD)	35.80 / 13.45 / 27.93	27.18 / 9.20 / 22.04	23.88 / 8.03 / 19.30	28.59 / 8.94 / 22.95	28.86 / 9.90 / 23.07
	D^2TV (V-KD)	34.60 / 12.70 / 26.86	27.75 / 9.71 / 22.63	24.36 / 8.00 / 19.41	31.53 / 10.28 / 24.83	29.56 / 10.17 / 23.43
	D^2TV	36.12 / 13.21 / 27.99	28.87 / 10.26 / 23.77	25.53 / 8.69 / 20.72	32.56 / 10.73 / 25.71	30.77 / 10.72 / 24.53
	LCMHA (ours)	**37.64 / 14.30 / 29.31**	**30.35 / 10.85 / 24.69**	**26.75 / 8.82 / 21.51**	**32.33 / 10.87 / 25.91**	**31.77 / 11.21 / 25.35**
	D^2TV* (w/o KD)	35.03 / 12.50 / 27.17	22.97 / 7.37 / 18.65	23.50 / 7.81 / 18.95	27.14 / 8.04 / 21.17	27.16 / 8.93 / 21.49
	D^2TV* (V-KD)	34.84 / 12.52 / 26.98	24.29 / 7.76 / 19.81	24.49 / 7.80 / 19.64	29.06 / 8.83 / 22.84	28.17 / 9.22 / 22.31
	D^2TV*	34.78 / 12.36 / 26.81	26.13 / 8.39 / 21.15	24.84 / 8.28 / 20.06	28.60 / 8.44 / 22.30	28.59 / 9.37 / 22.58
	LCMHA* (ours)	**35.56 / 12.99 / 27.55**	**28.96 / 9.89 / 23.69**	**25.71 / 8.40 / 20.70**	**32.18 / 10.29 / 25.32**	**30.60 / 10.39 / 24.32**
	LLaMA	42.89 / 19.34 / **33.98**	32.92 / **12.46** / 26.85	30.28 / 11.43 / 24.36	32.16 / 10.67 / 25.31	34.57 / 13.48 / 27.63
	LCMHA† (text only)	42.24 / 18.51 / 33.01	32.64 / 11.81 / 26.24	**31.69** / 11.83 / **24.89**	**34.15 / 11.77 / 26.85**	35.18 / **13.49** / 27.75
	LCMHA† (ours)	**43.46 / 19.47** / 33.95	**33.16** / 11.93 / **26.94**	31.40 / **11.87** / 24.71	33.10 / 10.42 / 25.91	**35.29**/ 13.43 / **27.88**
Indonesian →	VG-mT5	33.29 / 11.30 / 25.68	33.65 / 13.59 / 27.46	25.72 / 8.95 / 20.84	32.86 / 11.28 / 26.67	31.38 / 11.28 / 25.16
	D^2TV (w/o KD)	32.59 / 11.67 / 25.42	34.43 / 14.56 / 28.43	24.38 / 8.70 / 20.01	30.65 / 10.30 / 24.95	30.51 / 11.31 / 24.70
	D^2TV (V-KD)	32.88 / 11.56 / 25.45	32.67 / 13.01 / 26.71	25.50 / 8.97 / 20.65	32.48 / 11.31 / 25.88	30.88 / 11.21 / 24.67
	D^2TV	34.54 / 12.10 / 26.50	33.94 / 14.08 / 28.05	26.40 / 9.27 / 21.35	33.45 / 11.38 / 26.60	32.08 / 11.71 / 25.63
	LCMHA (ours)	**35.30 / 12.69 / 27.18**	**35.83 / 14.99 / 29.52**	**27.37 / 9.38 / 22.01**	**34.28 / 12.13 / 27.65**	**33.20 / 12.30 / 26.59**
	D^2TV* (w/o KD)	33.87 / 11.51 / 26.14	29.81 / 11.33 / 24.29	24.26 / 8.40 / 19.65	29.05 / 8.99 / 22.87	29.25 / 10.06 / 23.24
	D^2TV* (V-KD)	33.68 / 11.68 / 25.80	30.49 / 11.58 / 24.84	24.22 / 8.28 / 19.48	29.16 / 8.77 / 23.05	29.38 / 10.07 / 23.29
	D^2TV*	34.18 / 11.75 / 26.17	31.25 / 11.75 / 25.30	24.99 / 8.66 / 20.29	29.56 / 8.88 / 23.18	30.00 / 10.26 / 23.74
	LCMHA* (ours)	**34.98 / 12.42 / 27.01**	**32.76 / 12.77 / 26.75**	**27.10 / 9.41 / 21.85**	**32.77 / 10.44 / 25.87**	**31.90 / 11.26 / 25.37**
	LLaMA	40.17 / **16.97**/ 31.16	37.14 / **15.95** / 30.66	30.75 / 11.93 / 25.23	33.59 / **12.05** / 27.15	35.42 / **14.23 / 28.55**
	LCMHA† (text only)	39.02 / 15.93 / 30.22	37.12 / 15.50 / 30.23	**31.57 / 12.32 / 25.31**	**34.49** / 11.88 / **27.63**	35.55 / 13.91 / 28.35
	LCMHA† (ours)	**40.57** / 16.55 / **31.17**	**37.70** / 15.65 / **30.89**	31.36 / 11.93 / 24.84	33.66 / 10.87 / 26.68	**35.83** / 13.76 / 28.40
Russian →	VG-mT5	31.89 / 10.71 / 24.71	27.65 / 9.73 / 22.82	29.42 / 11.35 / 23.92	32.14 / 10.64 / 26.17	30.27 / 10.61 / 24.41
	D^2TV (w/o KD)	30.79 / 9.82 / 24.02	27.37 / 9.84 / 22.70	29.67 / 11.67 / 24.33	30.53 / 10.04 / 24.92	29.59 / 10.37 / 24.09
	D^2TV (V-KD)	31.18 / 10.64 / 24.26	26.50 / 9.60 / 21.91	28.32 / 10.93 / 23.00	31.38 / 10.76 / 25.25	29.34 / 10.48 / 23.60
	D^2TV	32.87 / 11.06 / 25.59	29.03 / 10.59 / 23.64	29.90 / 11.44 / 24.75	**33.29 / 11.88 / 26.96**	31.27 / 11.24 / 25.13
	LCMHA (ours)	**33.29 / 11.48 / 26.22**	**29.56 / 10.61 / 24.11**	**31.83 / 12.78 / 25.85**	33.08 / 11.05 / 26.66	**31.94 / 11.48 / 25.71**
	D^2TV* (w/o KD)	32.94 / 11.40 / 25.74	24.58 / 8.43 / 20.09	28.10 / 10.37 / 22.66	27.44 / 8.54 / 21.59	28.27 / 9.68 / 22.52
	D^2TV* (V-KD)	32.86 / 11.54 / 25.64	24.63 / 8.40 / 20.15	28.27 / 10.31 / 22.64	28.50 / 8.87 / 22.56	28.56 / 9.78 / 22.75
	D^2TV*	33.61 / 11.65 / 26.09	26.57 / 8.96 / 21.63	28.39 / 10.47 / 22.92	28.39 / 8.81 / 22.68	29.24 / 9.97 / 23.33
	LCMHA* (ours)	**33.82 / 12.05 / 26.50**	**28.96 / 9.86 / 23.54**	**30.04 / 11.53 / 24.26**	**32.25 / 10.25 / 25.77**	**31.27 / 10.92 / 25.02**
	LLaMA	38.32 / 15.58 / 29.75	30.73 / 11.36 / 25.25	34.18 / 14.31 / 27.54	32.40 / 10.83 / 26.51	33.91 / 13.02 / 27.27
	LCMHA† (text only)	37.54 / 15.00 / 29.25	31.53 / 11.68 / 25.97	35.36 / 14.67 / 28.39	**34.12 / 11.18 / 27.00**	34.64 / 13.14 / 27.66
	LCMHA† (ours)	**39.32 / 15.84 / 30.29**	**32.07 / 11.86 / 26.13**	**35.50 / 14.97 / 28.75**	32.44 / 10.16 / 25.59	**34.84 / 13.21 / 27.69**
Urdu →	VG-mT5	29.82 / 9.45 / 23.29	27.49 / 9.90 / 22.73	23.17 / 7.42 / 18.94	37.59 / 15.41 / 30.41	29.52 / 10.55 / 23.84
	D^2TV (w/o KD)	28.94 / 9.29 / 22.81	26.43 / 8.84 / 21.74	20.47 / 6.44 / 16.74	37.72 / 15.78 / 30.97	28.39 / 10.17 / 23.14
	D^2TV (V-KD)	29.60 / 9.63 / 23.00	26.30 / 9.02 / 21.65	22.58 / 7.36 / 18.03	37.52 / 15.25 / 30.19	29.00 / 10.31 / 23.21
	D^2TV	32.01 / 9.99 / 24.71	28.23 / 10.01 / 23.19	24.52 / 7.87 / 19.98	38.05 / 16.12 / 31.30	30.70 / 10.91 / 24.71
	LCMHA (ours)	**32.37 / 10.36 / 25.40**	**28.63 / 10.33 / 23.50**	**25.14 / 7.98 / 20.35**	**39.74 / 17.17 / 32.35**	**31.47 / 11.46 / 25.40**
	D^2TV* (w/o KD)	31.54 / 10.42 / 24.53	22.94 / 7.62 / 18.88	22.32 / 7.13 / 18.32	35.86 / 13.49 / 28.48	28.17 / 9.66 / 22.55
	D^2TV* (V-KD)	30.96 / 10.20 / 24.34	23.72 / 8.16 / 19.49	21.94 / 6.79 / 17.90	35.83 / 13.53 / 28.39	28.11 / 9.67 / 22.53
	D^2TV*	31.65 / 10.63 / 24.90	25.47 / 8.52 / 20.68	22.38 / 7.19 / 18.44	36.46 / 13.76 / 28.75	28.99 / 10.03 / 23.19
	LCMHA* (ours)	**32.69 / 10.72 / 25.00**	**27.99 / 10.04 / 23.27**	**24.64 / 7.92 / 19.95**	**37.61 / 15.44 / 30.20**	**30.73 / 11.03 / 24.60**
	LLaMA	35.66 / 13.78 / 27.86	28.85 / 10.22 / 23.72	26.96 / 9.72 / 22.23	34.92 / 15.11 / 28.77	31.60 / 12.21 / 25.65
	LCMHA† (text only)	37.16 / 13.78 / 28.36	30.84 / **11.27** / 25.38	**29.43 / 10.44 / 23.75**	39.72 / **16.62 / 31.84**	34.29 / 13.03 / 27.34
	LCMHA† (ours)	**38.49 / 15.09 / 29.97**	**31.59** / 11.24 / **26.11**	28.93 / 10.20 / 23.38	**39.89** / 16.59 / 31.68	**34.73 / 13.28 / 27.79**

Table 2: Performance comparison of different models in terms of ROUGE-1 / ROUGE-2 / ROUGE-L scores. [Model]* and [Model]† refers to the replacement of backbone from mT5 to mBART and LLaMA, respectively.

4.5 Ablation Analysis

The following ablation experiments are constructed to verify the effect of each component in the LCMHA model: a. Removing the visual input and the joint multimodal adapter, leaving only the text modality; b. Removing the language regularization term in LRH; c. Removing the language hyper adapters via LRH. As can be seen in Table 3, the ablation components have positive gains to the backbone model to varying degrees. Specifically, the performance degradation caused by removing visual input demonstrates the role of multimodal information. The adapter components improve performance by relaxing language parameter sharing to alleviate potential inter-language interference. LRH playing a role in promoting the learning for low-resource languages as it facilitates knowledge transfer in language-specific parameters.

4.6 Hypernetwork Dimension Analysis

This section analyzes the impact of the hypernetwork's hidden state dimensions h, as hidden state size primarily determines the overall size of LCMHA. We experimented with three configurations: LCMHA (base) with $h = 256$ dimensions, LCMHA (small) with $h = 128$ dimen-

	Models	English → *	Indonesian → *	Russian → *	Urdu → *
	LCMHA	31.77 / 11.21 / 25.35	33.20 / 12.30 / 26.59	31.94 / 11.48 / 25.71	31.47 / 11.46 / 25.40
a	*w/o* Multimodal	31.34 / 10.95 / 25.11	32.63 / 12.04 / 26.24	31.42 / 11.28 / 25.47	30.77 / 11.04 / 24.92
b	*w/o* Lang. Reg.	31.53 / 11.12 / 25.08	32.84 / 12.15 / 26.30	31.54 / 11.22 / 25.47	30.98 / 11.22 / 25.13
c	*w/o* HyperA.	31.12 / 10.74 / 24.75	32.66 / 11.89 / 26.11	31.17 / 10.94 / 25.19	30.40 / 10.71 / 24.60

Table 3: Ablation study of LCMHA based on the mT5 backbone in terms of average ROUGE-1 / ROUGE-2 / ROUGE-L scores.

Models	Extra Params	English → *	Indonesian → *	Russian → *	Urdu → *
LCMHA (base)	124M	31.77 / 11.21 / 25.35	33.20 / 12.30 / 26.59	31.94 / 11.48 / 25.71	31.47 / 11.46 / 25.40
LCMHA (small)	73M	31.82 / 11.23 / 25.27	32.96 / 12.30 / 26.41	31.90 / 11.63 / 25.88	31.23 / 11.44 / 25.19
LCMHA (tiny)	48M	31.78 / 11.15 / 25.30	32.98 / 12.21 / 26.39	31.82 / 11.43 / 25.56	31.38 / 11.48 / 25.35

Table 4: LCMHA size analysis based on the mT5 backbone in terms of average ROUGE-1 / ROUGE-2 / ROUGE-L scores.

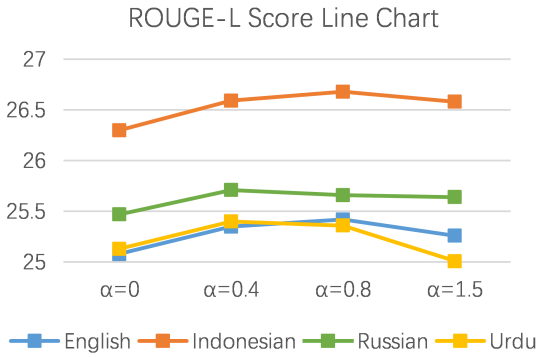


Figure 3: Average ROUGE-L score variation across languages under different values of α.

sions, and LCMHA (tiny) with $h = 64$ dimensions. The model performance and parameter counts are shown in Table 4. Generally, larger hypernetwork dimensions tend to yield better performance, although the differences are not obvious. We hypothesize that the limited number of language pairs used in our experiments might contribute to this observation.

4.7 Hyperparameter Analysis

This section further analyzes the impact of hyperparameter in the LRH by adjusting the weight α of the language regularization term, where $\alpha = 0$ means language regularization is not applied. The performance variations of the average ROUGE-L score are illustrated in Figure 3. We evaluated LCMHA with the mT5 backbone. Overall, LCMHA achieved the best results with α values of 0.4 or 0.8, while performance slightly declined when α was increased to 1.5.

Models	Consistency	Relevance	Coverage	Fluency
LLaMA	49.48	48.86	46.96	49.53
LCMHA†	**50.52**	**51.14**	**53.04**	**50.47**

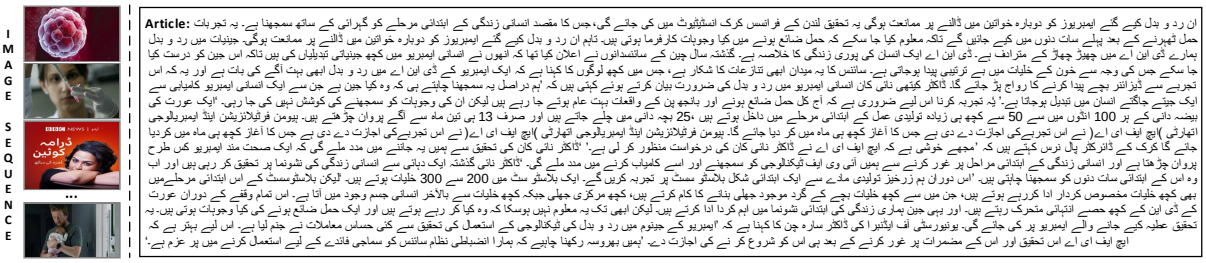
Table 5: Average win rates of LCMHA† and LLaMA from GPT-based evaluations.

4.8 GPT-4 Evaluation Results

Table 5 reports the average win rates from GPT-4 evaluations for summaries generated by LLaMA (LoRA) and LCMHA (on the LLaMA backbone), across four aspects: consistency, relevance, coverage, and fluency. The experimental results show that, while achieving comparable fluency and consistency, the proposed method demonstrates an advantage in relevance, and coverage, indicating that LCMHA better focuses on the key points of source documents and aligns with reference summaries in different languages. Complete evaluation results for each language direction are provided in Appendix B, where LCMHA obtained a clear winning rate in the Urdu direction, indicating the role of the proposed method in protecting and promoting the learning of low-resource language combinations.

4.9 Case Study

We present a case study in the Urdu directions using LLaMA as the backbone model, where LCMHA demonstrates more noticeable improvements in objective evaluation metrics. As shown in Figure 4, all models are able to generate reasonably good summaries that reflect different aspects of the source document. In some cases, LCMHA's outputs appear more closely aligned with the reference summaries. For instance, in the Urdu-to-Urdu di-



English Summary	**Reference:** UK scientists have been given the go-ahead by the fertility regulator to genetically modify human embryos. **LLaMA (LoRA):** Human embryos will be edited in UK research to understand why so many pregnancies end in miscarriage, a leading scientist has told the BBC. **LLaMA (LCMHA):** UK scientists have been given legal approval to edit the DNA of human embryos for research purposes.
Indonesian Summary	**Reference:** Para ilmuwan Inggris diizinkan pejabat pemerintah yang mengawasi masalah kesuburan untuk merekayasa gen janin manusia. **LLaMA (LoRA):** Seorang ilmuwan Inggris diberikan izin untuk melakukan perubahan pada DNA embrio manusia untuk pertama kalinya. **LLaMA (LCMHA):** Seorang ilmuwan Inggris akan melakukan percobaan pertama di Inggris untuk mengubah genetika embrio manusia.
Russian Summary	**Reference:** Британские ученые получили добро от управления по оплодотворению и эмбриологии человека для начала работ по генетической модификации человеческих эмбрионов. **LLaMA (LoRA):** Британский ученый получил разрешение на проведение эксперимента с человеческим эмбрионом, чтобы понять, как в раннем возрасте формируется человеческий геном. **LLaMA (LCMHA):** В Лондоне начато первое в мире клиническое исследование, в котором используются измененные ДНК человеческих эмбрионов.
Urdu Summary	**Reference:** برطانیہ کے سائنسدانوں کو انسانی ایمبریو)رحم مادر میں ناممکمل بچے(میں جینیاتی رد و بدل کی تحقیق کی اجازت مل گئی ہے۔ **LLaMA (LoRA):** ایک ڈاکٹر کو اجازت مل گئی ہے کہ، وہ حمل ضائع ہونے کے 30 سے 50 فیصد واقعات کی وجو **LLaMA (LCMHA):** برطانیہ میں سائنسدانوں نے انسانی ایمبریوز میں جینیاتی تبدیلیاں کرنے کے تجربے کے لیے اجازت حاصل کر لی ہے۔

Figure 4: M3S generation results of LLaMA (LoRA) and LLaMA (LCMHA) models on four directions: Urdu-English, Urdu-Indonesian, Urdu-Russian, and Urdu-Urdu.

rection, LLaMA (LoRA) generates: "A doctor has been granted permission to investigate the causes behind miscarriages," whereas LLaMA (LCMHA) produces: "Scientists in the UK have obtained permission to conduct experiments involving genetic modification of human embryos." This intuitively reflects the potential of LCMHA to better utilize high-resource knowledge, especially in directions like Urdu where LLaMA's exposure may be comparatively lower.

5 Conclusion

In this work, we introduce the language constrained multimodal hyper adapter (LCMHA) for many-to-many multimodal summarization (M3S). LCMHA incorporates language-specific multimodal adapters through a language regularized hypernetwork to maintain shared multimodal summarization knowledge while relaxing language-specific knowledge sharing. This approach enhances the adaptability of M3S models to multimodal summarization across various language combinations. Extensive experiments on the M3S benchmark demonstrate that the proposed method outperforms existing state-of-the-art approaches and confirms its scalability across multiple multilingual pre-trained backbones.

Limitations

Although LCMHA demonstrates improvements in M3S across different PLM backbones, this study only explores four languages and 16 language directions. Future work should extend the evaluation to a broader range of languages to further validate the applicability of the proposed method. Additionally, considering that LoRA is a common and efficient parameter-efficient fine-tuning approach for LLMs, adapting LCMHA to LoRA remains an interesting topic for further investigation, such as developing multimodal hyper-LoRA structures tailored for M3S.

Moreover, this work does not explicitly address broader societal and ethical risks in multilingual multimodal summarization. Knowledge transfer across languages may amplify existing biases, particularly for underrepresented languages, due to data imbalance or cultural semantic misalignment. Hallucinations in multimodal summaries can also mislead users or spread misinformation. While these issues are beyond the scope of this study, we acknowledge their significance and encourage future work to incorporate safeguards against hallucination, cultural misalignment, and bias amplification.

Acknowledgements

The work is supported by the National Natural Science Foundation of China under Grant 62406223, 62302334, Natural Science Foundation of Tianjin under Grant 24JCZDJC00130, and research funding from Cangzhou Institute of Tiangong University under Grant TGCYY-Z-0303.

References

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. Mad-g: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781.

Ankur Bapna and Orhan Firat. 2019. Simple, scalable adaptation for neural machine translation. 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 1538–1548.

Christos Baziotis, Mikel Artetxe, James Cross, and Shruti Bhosale. 2022. Multilingual machine translation with hyper-adapters. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1170–1185.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Yuan Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2023. Crosssum: Beyond english-centric cross-lingual summarization for 1,500+ language pairs. In *Annual Meeting of the Association of Computational Linguistics 2023*, pages 2541–2564. Association for Computational Linguistics (ACL).

Jingqiang Chen and Hai Zhuge. 2018. Abstractive text-image summarization using multi-modal attentional hierarchical rnn. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4046–4056.

Cunhang Fan, Wang Xiang, Jianhua Tao, Jiangyan Yi, and Zhao Lv. 2025. Cross-modal knowledge distillation with multi-stage adaptive feature fusion for speech separation. *IEEE Transactions on Audio, Speech and Language Processing*.

Xiachong Feng, Xiaocheng Feng, Xiyuan Du, Min-Yen Kan, and Bing Qin. 2024. Adapter-based selective knowledge distillation for federated multi-domain meeting summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Xiyan Fu, Jun Wang, and Zhenglu Yang. 2021. Mm-avs: A full-scale dataset for multi-modal summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5922–5926.

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Gaurav Pandey, Dinesh Raghu, and Setu Sinha. 2024. Healthalignsumm: Utilizing alignment for multimodal summarization of code-mixed healthcare dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11546–11560.

David Ha, Andrew M Dai, and Quoc V Le. 2022. Hypernetworks. In *International Conference on Learning Representations*.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Bo He, Jun Wang, Jielin Qiu, Trung Bui, Abhinav Shrivastava, and Zhaowen Wang. 2023. Align and attend: Multimodal summarization with dual contrastive losses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14867–14878.

Tsutomu Hirao, Naoki Kobayashi, Hidetaka Kamigaito, Manabu Okumura, and Akisato Kimura. 2024. Video discourse parsing and its application to multimodal summarization: A dataset and baseline approaches. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9943–9958.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

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*.

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Lee. 2023. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5276.

Yichong Huang, Xiaocheng Feng, Xinwei Geng, Baohang Li, and Bing Qin. 2023. Towards higher pareto frontier in multilingual machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3802–3818.

Jinbae Im, Moonki Kim, Hoyeop Lee, Hyunsouk Cho, and Sehee Chung. 2021. Self-supervised multimodal opinion summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 388–403.

Hamish Ivison, Akshita Bhagia, Yizhong Wang, Hannaneh Hajishirzi, and Matthew E Peters. 2023. Hint: Hypernetwork instruction tuning for efficient zero- and few-shot generalisation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11272–11288.

Anubhav Jangra, Sourajit Mukherjee, Adam Jatowt, Sriparna Saha, and Mohammad Hasanuzzaman. 2023.

A survey on multi-modal summarization. *ACM Computing Surveys*, 55(13s):1–36.

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

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.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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.

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

Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2018. Read, watch, listen, and summarize: Multi-modal summarization for asynchronous text, image, audio and video. *IEEE Transactions on Knowledge and Data Engineering*, 31(5):996–1009.

Manling Li, Lingyu Zhang, Heng Ji, and Richard J Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196.

Mingzhe Li, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2020. Vmsmo: Learning to generate multimodal summary for video-based news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9360–9369.

Yunlong Liang, Fandong Meng, Jiaan Wang, Jinan Xu, Yufeng Chen, and Jie Zhou. 2023a. D2tv: Dual knowledge distillation and target-oriented vision modeling for many-to-many multimodal summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14910–14922.

Yunlong Liang, Fandong Meng, Jinan Xu, Jiaan Wang, Yufeng Chen, and Jie Zhou. 2023b. Summary-oriented vision modeling for multimodal abstractive summarization. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

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

Dengtian Lin, Liqiang Jing, Xuemeng Song, Meng Liu, Teng Sun, and Liqiang Nie. 2023. Adapting generative pretrained language model for open-domain multimodal sentence summarization. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 195–204.

Nayu Liu, Xian Sun, Hongfeng Yu, Fanglong Yao, Guangluan Xu, and Kun Fu. 2022a. Abstractive summarization for video: A revisit in multistage fusion network with forget gate. *IEEE Transactions on Multimedia*, 25:3296–3310.

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.

Nayu Liu, Kaiwen Wei, Xian Sun, Hongfeng Yu, Fanglong Yao, Li Jin, Guo Zhi, and Guangluan Xu. 2022b. Assist non-native viewers: Multimodal cross-lingual summarization for how2 videos. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6959–6969.

Nayu Liu, Kaiwen Wei, Yong Yang, Jianhua Tao, Xian Sun, Fanglong Yao, Hongfeng Yu, Li Jin, Zhao Lv, and Cunhang Fan. 2024. Multimodal cross-lingual summarization for videos: A revisit in knowledge distillation induced triple-stage training method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Bo Lv, Xin Liu, Kaiwen Wei, Ping Luo, and Yue Yu. 2024. TAeKD: Teacher assistant enhanced knowledge distillation for closed-source multilingual neural machine translation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15530–15541.

Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. 2021. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 565–576.

Louis Mahon and Mirella Lapata. 2024. A modular approach for multimodal summarization of tv shows. *arXiv preprint arXiv:2403.03823*.

Laura Nguyen, Thomas Scialom, Benjamin Piwowarski, and Jacopo Staiano. 2023. Loralay: A multilingual and multimodal dataset for long range and layout-aware summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 636–651.

Keighley Overbay, Jaewoo Ahn, Joonsuk Park, Gunhee Kim, et al. 2023. mredditsum: A multimodal abstractive summarization dataset of reddit threads with images. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4117–4132.

Shruti Palaskar, Jindřich Libovickỳ, Spandana Gella, and Florian Metze. 2019. Multimodal abstractive summarization for how2 videos. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6587–6596.

Vaidehi Patil, Leonardo Ribeiro, Mengwen Liu, Mohit Bansal, and Markus Dreyer. 2024. Refinesumm: Self-refining mllm for generating a multimodal summarization dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13773–13786.

Jielin Qiu, Jiacheng Zhu, Mengdi Xu, Franck Dernoncourt, Trung Bui, Zhaowen Wang, Bo Li, Ding Zhao, and Hailin Jin. 2022. Mhms: Multimodal hierarchical multimedia summarization. *arXiv preprint arXiv:2204.03734*.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Xindi Shang, Zehuan Yuan, Anran Wang, and Changhu Wang. 2021. Multimodal video summarization via time-aware transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1756–1765.

Xiaorui Shi. 2024. Towards making the most of knowledge across languages for multimodal cross-lingual summarization. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 424–438. Springer.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. 2022. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5227–5237.

Peggy Tang, Kun Hu, Lei Zhang, Jiebo Luo, and Zhiyong Wang. 2023. Tldw: Extreme multimodal summarisation of news videos. *IEEE Transactions on Circuits and Systems for Video Technology*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466.

Yi Tay, Zhe Zhao, Dara Bahri, Donald Metzler, and Da-Cheng Juan. 2020. Hypergrid transformers: Towards a single model for multiple tasks. In *International conference on learning representations*.

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

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, Gertjan van Noord, and Sebastian Ruder. 2022. Hyper-x: A unified hypernetwork for multi-task multilingual transfer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7934–7949. Association for Computational Linguistics (ACL).

Yash Verma, Anubhav Jangra, Raghvendra Verma, and Sriparna Saha. 2023. Large scale multi-lingual multimodal summarization dataset. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3620–3632.

Johannes Von Oswald, Christian Henning, Benjamin F Grewe, and João Sacramento. 2020. Continual learning with hypernetworks. In *8th International Conference on Learning Representations*. International Conference on Learning Representations.

Jiaan Wang, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023. Zero-shot cross-lingual summarization via large language models. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 12–23.

Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022. A survey on cross-lingual summarization. *Transactions of the Association for Computational Linguistics*, 10:1304–1323.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuan-Jing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021. K-adapter: Infusing knowledge into pre-trained models with adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418.

Xinyi Wang, Yulia Tsvetkov, and Graham Neubig. 2020. Balancing training for multilingual neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8526–8537.

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*.

Min Xiao, Junnan Zhu, Feifei Zhai, Yu Zhou, and Chengqing Zong. 2024. Diusum: Dynamic image

utilization for multimodal summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19297–19305.

Min Xiao, Junnan Zhu, Feifei Zhai, Chengqing Zong, and Yu Zhou. 2025. Pay more attention to images: Numerous images-oriented multimodal summarization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9379–9392.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

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.

Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021. Vision guided generative pre-trained language models for multimodal abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3995–4007.

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

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.

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

Hao Zhao, Zihan Qiu, Huijia Wu, Zili Wang, Zhaofeng He, and Jie Fu. 2024. Hypermoe: Towards better mixture of experts via transferring among experts. *arXiv preprint arXiv:2402.12656*.

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, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2021. Graph-based multimodal ranking models for multimodal summarization. *Transactions on Asian and low-resource language information processing*, 20(4):1–21.

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

Haojie Zhuang, Wei Emma Zhang, Leon Xie, Weitong Chen, Jian Yang, and Quan Sheng. 2024. Automatic, meta and human evaluation for multimodal summarization with multimodal output. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7761–7783.

Hyperparameters	**Backbone**		
	mT5	**mBART**	**LLaMA**
Max length for article	520	520	1024
Max length for summary	84	84	128
Max length for visual feat. seq.	108	108	108
Batch size for per GPU	8	8	4
Gradient accumulation	4	4	4
Learning rate	8e-5	3e-5	1e-4
Warmup ratio	0.05	0.05	0.1
Training epochs	15	15	3
Optimizer	Adam	Adam	Adam
Adam beta1	0.9	0.9	0.9
Adam beta2	0.998	0.98	0.98
GPU model	A100×1	A100×1	A100×4
Avg. num of algorithm runs	2	2	2
Beam search size	3	3	1
Length penalty	1	1	1
Hidden size	768	1024	4096
Num of layers	12+12	12+12	32
Attention heads	12	16	32
Num of hypernetwork layers l	2	2	2
Hypernetwork hidden size h	256	256	64
Source language ID dim d_e	50	50	50
Target language ID dim d_e	50	50	50
Layer ID dim d_e	50	50	50
Language regularization α	0.4	0.4	0.4
Adapter bottleneck dim d_b	256	256	256

Table 6: Experimental settings for LCMHA.

A Implementation Details.

Table 6 shows the complete experimental settings regarding data processing, training, and inference. Following (Liang et al., 2023a,b; Bhattacharjee et al., 2023), the maximum sequence length for input articles and summaries is truncated to 512 and 84, respectively; Image features are extracted via a pretrained Faster R-CNN object detector, with

Models	English → *	Indonesian → *	Russian → *	Urdu → *
	Cons. / Rel. / Cov. / Flu.	Cons. / Rel. / Cov. / Flu.	Cons. / Rel. / Cov. / Flu.	Cons. / Rel. / Cov. / Flu.
LLaMA	49.90 / 48.78 / 47.06 / 50.00	51.63 / 50.33 / 49.60 / 49.87	49.79 / 49.01 / 47.11 / 49.41	46.47 / 47.18 / 43.93 / 48.84
LCMHA†	50.10 / 51.22 / 52.94 / 50.00	48.37 / 49.67 / 50.40 / 50.13	50.21 / 50.99 / 52.89 / 50.59	53.53 / 52.82 / 56.07 / 51.16

Table 7: Win rates in each language direction from GPT-4 evaluations, including consistency, relevance, coverage, and fluency.

the maximum image sequence length truncated to 108. The proposed LCMHA is integrated with two multilingual pretrained models, mT5[6] (Xue et al., 2021) and mBART[7] (Tang et al., 2021) as used in previous works, and an LLM backbone, LLaMA-3-8B[8] (Touvron et al., 2023). The prompt used for LLaMA is as follows:

Summarize the following {src_lang} text into a {tgt_lang} abstract:

The number of hypernetwork layers is $l = 2$, with the hidden state dimension $h = 256$ for mT5 and mBART and $h = 64$ for LLaMA, and the source language, target language, and layer ids dimensions are all $d_e = 50$. The hyperparameter in LRH is set to $\alpha = 0.4$. The adapter bottleneck dimension is $d_b = 256$.

For training, the models are optimized using Adam (Kingma and Ba, 2014) with an initial learning rate of 8e-5 for mT5, 3e-5 for mBART, and 1e-4 for LLaMA backbones respectively. Specifically, due to the high memory and computational costs of LLaMA, we use LoRA to train LCMHA and the backbone. We train 15 epochs with a batch size of 8 and gradient accumulation set to 4 for mT5 and mBART on an A100 GPU, and 3 epochs with a batch size of 16 and gradient accumulation set to 4 for LLaMA on 4 A100 GPUs. For inference, we use beam search with a beam size of 3 for mT5 and mBART, and use nucleus sampling (Holtzman et al., 2020) with p=0.9 for LLaMA.

B GPT-4 Evaluation

We use GPT-4 to evaluate generated summaries, including consistency, relevance, coverage, and fluency, where the template is provided in Table 5. For each language combination, we test 100 samples. Specially, To mitigate the potential influence of method placement order on GPT-4 evaluations, each baseline and LCMHA are evaluated twice: once as [Method A] vs. [Method B], and then with the order reversed. The final score is obtained by averaging the results from both evaluations.

Table 7 lists the experimental results of LLaMA (LoRA) and LCMHA (on the LLaMA backbone) in each language direction, corresponding to the average GPT-4 evaluation scores for all language combinations in Section 4.8 of the main paper content.

Prompt for GPT-4 Evaluation

You are an expert evaluator in summarization. Your task is to compare two machine-generated summaries based on a given {source_lang} article and a {tgt_lang} reference summary.

Evaluation Criteria:
Compare the two summaries for each of the following aspects:
1. Consistency: Which summary more accurately reflects facts from the source document without introducing hallucinations?
2. Relevance: Which summary focuses better on the key points of the source document while avoiding irrelevant content?
3. Coverage: Which summary better includes the important aspects of the reference summary while maintaining informativeness?
4. Fluency: Which summary is more grammatically correct, well-structured, and easy to read?

Article:
{article}
Reference Summary:
{reference summary}
Method A Summary:
{method A summary}
Method B Summary:
{method B summary}

In the following comparison of the outputs of two models, which model produces better results in the evaluation criteria below? Only one superior model can be selected.
1. Consistency Comparison: <A > or <B>
2. Relevance Comparison: <A > or <B>
3. Coverage Comparison: <A > or <B>
4. Fluency Comparison: <A > or <B>

Only output the metrics and their corresponding comparison results.

Figure 5: GPT-4 templates for ranking generated summaries.

[6] https://huggingface.co/google/mt5-base
[7] https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt
[8] https://huggingface.co/meta-llama/Meta-Llama-3-8B