ViInfoSum: A Novel Dataset for Multimodal Summarization in Vietnamese

Nguyen Khoi Tran[1,2], Binh Xuan Le[1,2], Nghia Hieu Nguyen[1,2],
Khiem Vinh Tran[1,2], Ngan Luu-Thuy Nguyen[1,2],
[1]University of Information Technology,
[2]Vietnam National University, Ho Chi Minh city, Viet Nam
{22520987, 22520131}@gm.uit.edu.vn, {nghiangh, khiemvt, ngannlt}@uit.edu.vn
Correspondence: ngannlt@uit.edu.vn

Abstract

Multimodal Summarization requires the exploration of mutual information from multimodal data such as text and graphics. Although this task has been well-studied on high resource languages such as English, the development in methodology of this task on low-resource languages such as Vietnamese is limited. In this study, we introduce ViInfoSum, a novel dataset for evaluating Multimodal Summarization methods in Vietnamese. Evaluation of baselines for abstractive text summarization and OCR-based image captioning methods show that they are not effective in determining correct entities from multimodal inputs, which results in wrong summarization. To this end, we propose ViInfoSummer, a novel method that learn to indicate entities correctly from multimodal inputs which leads to performance improvement on the ViInfoSum dataset for Multimodal Summarization in Vietnamese.

1 Introduction

Infographics convey information through a combination of visual and textual elements. Summarizing such data therefore requires multimodal understanding and reasoning, which is substantially more challenging than traditional text-only summarization. Advances in this research area provide fundamental techniques for downstream applications such as human–machine interaction, visual analytics, and multimodal summarization.

Although this task has been extensively studied in high-resource languages such as English (He et al., 2023; Zhu et al., 2021, 2018, 2020), research on infographic summarization for low-resource languages, including Vietnamese, remains limited. Linguistic differences between English and Vietnamese introduce unique challenges, necessitating specialized techniques to effectively address this task in Vietnamese and other typologically similar languages. To this end, we introduce **ViInfoSum** (**Vi**etnamese **Info**graphics **Cap**tioning), a novel dataset for multimodal summarization in Vietnamese. Furthermore, we evaluate existing captioning and summarization methods on the proposed dataset. Experimental results show that these baseline approaches struggle to accurately identify the main topic or subject of Vietnamese infographics. To address this limitation, we propose *ViInfoSummer*, a novel model designed to more effectively recognize infographic topics and generate high-quality summaries in Vietnamese.

Our contributions can be summarized as twofold:

- We introduce *ViInfoSum*, a new dataset for Vietnamese infographic captioning and multimodal summarization, addressing the lack of resources for low-resource languages.
- We propose *ViInfoSummer*, a novel model that improves entity recognition for Vietnamese infographic summarization.

2 Related Works

2.1 Datasets

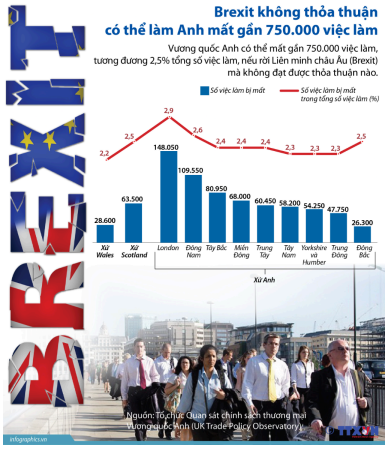
Early research on image captioning primarily focused on describing objects and natural visual contexts, with representative datasets such as MS COCO (Lin et al., 2014). While these datasets established a strong foundation for visual description, they exhibit inherent limitations when applied to text-rich images, as they lack explicit modeling of characters and numerical information. Subsequent datasets, such as TextCaps (Sidorov et al., 2020) and Screen2Word (Wang et al., 2021a), introduced OCR signals into the reasoning process. These datasets implicitly define a form of visual grammar based on non-linear spatial layouts, where information is conveyed through the interplay of text, charts, and graphical cues, which requires models to integrate textual reading with visual context for better caption and summarization generation.

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 339–350
October 17-21, 2026 ©2026 Association for Computational Linguistics



Summarization

chủ đề của ngày du lịch thế giới 27/9/2020 là du lịch và phát triển nông thôn . ngày này là cơ hội để nhìn nhận vai trò và khả năng kiến tạo tương lai của du lịch tại các khu vực nông thôn *(The theme of World Tourism Day on September 27, 2020 was **"Tourism and Rural Development."** This day provided an opportunity to reflect on the role of tourism and its potential to shape the future of rural areas)*



Summarization

vương quốc anh có thể mất gần 750.000 việc làm tương đương 25 tổng số việc làm nếu rời liên minh châu âu brexit mà không đạt được thoả thuận nào *(The United Kingdom could lose nearly 750,000 jobs, equivalent to 25% of total employment, if it leaves the European Union (Brexit) without reaching any agreement)*



Summarization

sea games 17 tổ chức tại singapore từ ngày 12 đến 20/6/1993 với 29 môn thi đấu chính thức . indonesia đứng đầu bảng tổng sắp với 253 huy chương *(The 17th SEA Games were held in Singapore from June 12 to June 20, 1993, featuring 29 official sports. Indonesia topped the medal table with 253 medals)*

Figure 1: Some examples from the ViInfoSum dataset.

2.2 Methods for Image Captioning

Early image captioning methods primarily focus on learning visual representations from images using region-level (Anderson et al., 2018) or grid-level features (Xu et al., 2015). While effective for natural image understanding, these approaches are limited in text-rich scenarios. To address this limitation, OCR-based image captioning methods explicitly incorporate text recognition as a core component of the modeling pipeline (Yang et al., 2021). A central challenge addressed by this paradigm is the high prevalence of out-of-vocabulary tokens, including numerical values, named entities, and domain-specific terminology. Inspired by the Pointer Network framework (Vinyals et al., 2015), (Hu et al., 2020; Singh et al., 2019) proposed novel methods to align OCR tokens with visual representations.

Subsequent studies further enhance multimodal reasoning by improving representation learning, pretraining strategies, and structural modeling. Anchor-Captioner (Xu et al., 2021) proposes a keyword-driven generation strategy that supports diverse semantic perspectives by conditioning captions on different textual anchors. In parallel, DEVICE (Xu et al., 2025) advances spatial modeling by capturing fine-grained and hierarchical layout structures within images, enabling deeper multimodal interactions. Despite these advances, existing methods remain largely focused on local entity grounding, limiting their capacity for global semantic aggregation and topic-level reasoning required for holistic infographic summarization.

2.3 Methods for Abstractive Text Summarization

For data with complex layouts and high text density, where conventional image captioning models often fail to capture the full semantic context, a promising direction is to extract textual content via OCR and generate descriptions using abstractive text summarization models. This paradigm is particularly suitable for text-rich visual data such as infographics. Recent advances in abstractive summarization have been largely driven by pretrained language models, including T5 (Raffel et al., 2020) and BART (Lewis et al., 2020), which achieve strong performance across multiple languages, including Vietnamese. Monolingual Vietnamese variants such as ViT5 (Phan et al., 2022) and BARTpho (Tran et al., 2022) further demonstrate substantial improvements over earlier approaches, enabling the generation of accurate, fluent, and linguistically well-formed summaries. However, these methods operate purely on linearized text and do not take into account the spatial structure of scene texts or the associated visual content, making them inadequate for holis-

tic infographic summarization, where layout and visual-textual interactions play a critical role.

3 ViInfoSum Dataset

3.1 Dataset Construction

We collected high-quality infographics in Vietnamese from the official website of the Vietnam News Agency (VNA)[1]. This website includes infographics whose visual and textual contents are in the wide range aspects: Events and Characters (EC), Science and Education (SE), National Defense and Security (NDS), Sports and Arts (SA), Disasters and Accidents (DA), Sociality and Culture (SC), Politics and Diplomacy (PD), Social Healthcare (SH), and Economy and Integration (EI).

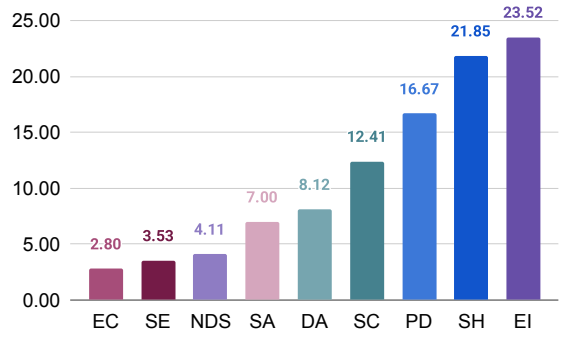


Figure 2: Distribution (%) of infographics in ViInfoSum according to the aspects.

Every infographic from the official website of VNA is formatted as PNG and is associated by a summarization written by the staffs of the VNA (Figure 1). These summarizations were written in news writing style. We preprocessed to remove from them the date, tag, and location of publication.

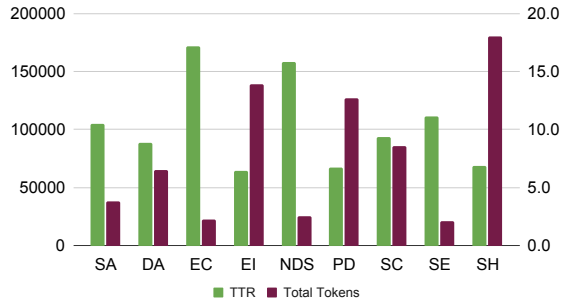


Figure 3: TTR and Total Tokens (%) according to the aspects.

[1]https://infographics.vn/

3.2 Overall Statistics

Statistically, the ViInfoSum includes **17,840** infographics, each of them has at most one summarization. The distribution of these infographics according to the aspects is given in the Figure 2. Most of the infographics in our dataset are relevant to Economy and Integration, Social Healthcare, and Politics and Diplomacy, while Education and Science, and Events and Characters occupy the smallest number. Finally, the dataset is then organized into three sets (train set, dev set, and test set) following the ratio of 8:1:1.

In addition, we defined the metric TTR (Typed Token Ratio) to determine the popularity of tokens in the summarization in our ViInfoSum dataset. In particular, TTR is calculated as:

$$TTS = \frac{\text{number of unique tokens}}{\text{number of tokens}} \quad (1)$$

According to Figure 3, in terms of corpus size, the Social Health (SH) domain dominates the dataset in total token count, followed by Economics and Integration (EI) and Politics and Diplomacy (PD), whereas domains such as National Security and Defense (NSD) and Science and Education (SE) are considerably smaller. Examining the TTR curve, we observe a clear inverse correlation between corpus size and TTR: domains with larger unique token volumes tend to exhibit lower TTR values. This phenomenon is consistent with well-established principles in statistical linguistics, whereby increasing corpus size leads to higher repetition of function words and domain-specific terminology, thereby reducing the proportion of newly introduced word types. Conversely, smaller domains such as Events and Characters (CE) achieve the highest TTR, largely due to the dense concentration of named entities and unique proper nouns.

To assess the degree of overlap between caption content and textual information directly appearing in images, we extract textual features using PaddleOCR (Cui et al., 2025) in combination with VietOCR[2] . The overlap ratio is defined as the proportion of tokens in a caption that also appear in the OCR-extracted text.

The statistics reported in Figure 4 reveal a strong correlation between caption content and OCR data. Specifically, more than 80% of captions contain at least 30% of their tokens overlapping with OCR text. Notably, 40.36% of samples exhibit near-

[2]https://github.com/pbcquoc/vietocr.git

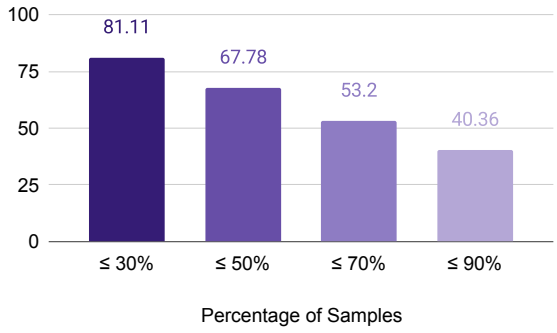


Figure 4: The degree of overlap (%) between tokens in the summarization and the OCR in the infographic.

verbatim overlap ($\geq 90\%$), indicating that reusing textual elements present in the image is a dominant pattern in this dataset. These findings suggest that, in the context of this dataset, captioning and summarization often rely heavily on lexical reuse rather than purely abstractive generation.

4 Methodology

We introduce ViInfoSummer which is a novel approach that enhances entity-level consistency by maximizing the similarity between predicted entities and their ground-truth counterparts, while simultaneously capturing critical visual layout information beyond the reliance on prominent textual features in the input (Figure 5).

4.1 Feature Extraction

Patch Features. We employ LayoutLMv3 to extract visual patch features $\{x_m^{\text{pat}}\}_{m=1}^{M}$ along with their corresponding bounding box coordinates $\{b_m^{\text{pat}}\}_{m=1}^{M}$.

Layout OCR Features. We adopt a three-stage pipeline for textual feature extraction from infographics: (i) layout detection, (ii) text detection, and (iii) text aggregation. Given an input image, layout regions $\{b_n^{\text{lay}}\}_{n=1}^{N}$ are detected using the PaddlePaddle PP-DocLayout-L model (Cui et al., 2025). For each detected layout, we apply VietOCR to extract the corresponding textual content $\{t_n^{\text{lay}}\}_{n=1}^{N}$. To obtain semantic representations of scene texts, we encode the extracted texts using PhoBERT (Nguyen and Nguyen, 2020), resulting in layout-level textual features $\{x_n^{\text{t-lay}}\}_{n=1}^{N}$.

The input source text S_{source} for the captioning model is constructed by concatenating all layout texts, ordered in a left-to-right, top-to-bottom manner to preserve the natural reading flow. Layout spatial features are modeled using 2D positional embeddings by transforming each layout bounding box b_n^{lay} into a normalized geometric representation $b_n^{\text{lay-trans}}$ as follows:

$$b_n^{\text{lay-trans}} = [x_{tl}; y_{tl}; x_{tr}; y_{tr}; x_{bl}; y_{bl}; x_{br}; y_{br}; w; h] \quad (2)$$

where (x_{tl}, y_{tl}), (x_{tr}, y_{tr}), (x_{bl}, y_{bl}), and (x_{br}, y_{br}) denote the coordinates of the four corners of the layout bounding box, and w, h represent its width and height.

Finally, the layout feature representation $\{x_n^{\text{lay}}\}_{n=1}^{N}$ is obtained by combining textual semantics with spatial information:

$$x_n^{\text{lay}} = x_n^{\text{t-lay}} + W_p \frac{\text{PosEmbedding}(b_n^{\text{lay-trans}})}{10} \quad (3)$$

where $\text{PosEmbedding}(\cdot)$ denotes a learnable 2D spatial embedding of size $224 \times d$, implemented via nn.Embedding, and $W_p \in \mathbb{R}^{d \times d}$ are learnable parameters.

4.2 Spatial Overlap Attention Module

Infographics typically contain dense textual content organized into multiple paragraphs that are spatially arranged to form distinct visual layouts. Treating OCR outputs as isolated single words, as commonly seen in traditional OCR-based image captioning approaches, has been shown to be ineffective for capturing such structured semantics. Rather than modeling relationships at the individual OCR token level, our approach emphasizes the semantic representation of textual paragraphs within each layout region, allowing the model to capture higher-level meaning across each layout, where the final caption relies heavily on the aggregated semantic content of OCR paragraphs and their layout-aware organization.

To determine which layout regions should be emphasized during caption generation, we design a cross-attention mechanism between visual patch features and layout features to identify layout importance. First, we compute the Intersection-over-Regions (IoRs) between a layout bounding box b_n^{lay} and b_m^{pat}, contains 3 main values: Intersection-over-union (IoU), Intersection-over-patch (IoP) and Intersection-over-layout (IoL). IoRs capture fine-grained 2D semantic positional relationships by quantifying how visual patches spatially intersect with layout regions.

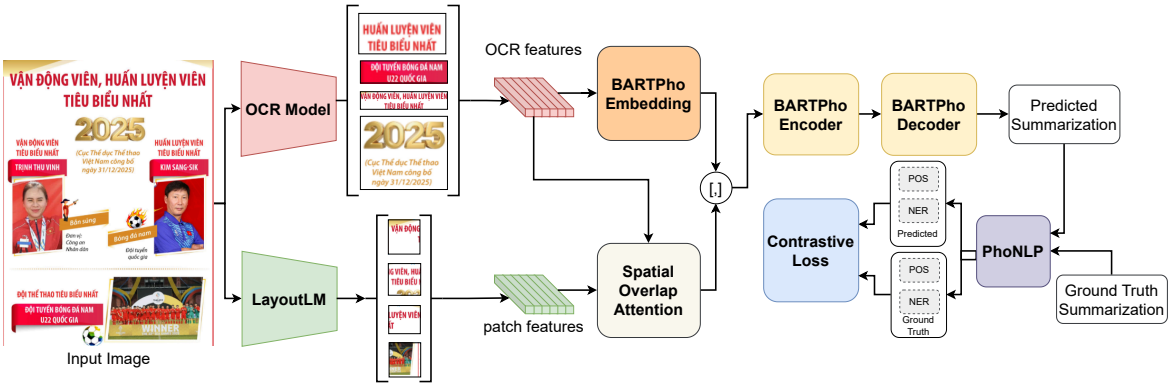


Figure 5: The overall architecture of the ViInfoSummer method.

$$\text{IoRs}_{mn} = \frac{\text{IoU}_{mn} + \text{IoP}_{mn} + \text{IoL}_{mn}}{3} \quad (4)$$

Inspired by Fang et al. (Fang et al., 2023), the intersection ratios are embedded into a lookup table implemented via an nn.Embedding layer of size 32×16, where 16 corresponds to the number of attention heads. Since $\text{IoRs}_{mn} \in [0, 1]$, we discretize the overlap score to map continuous overlap values into the discrete index space of the lookup table:

$$\phi = (\text{IoRs}_{mn} \times 31)\,.long() \quad (5)$$

To perform cross-attention mechanism, we calculate the query Q, key K, and value V as follows:

$$\begin{cases} Q = x^{pat} W_Q \\ K = x^{lay} W_K \\ V = x^{lay} W_V \end{cases} \quad (6)$$

where query $Q \in \mathbb{R}^{M \times d}$ and key K, value $V \in \mathbb{R}^{N \times d}$. Then, we can obtain the attention weights matrix A by:

$$A = \frac{QK^T}{\sqrt{d}}, \quad (7)$$

where $A \in \mathbb{R}^{M \times N}$. Moreover, to mitigate the influence of outlier patches that are spatially distant from any layout region, we introduce a Euclid Distance Matrix D that encodes the geometric distance between b_m^{pat} and b_n^{lay} by:

$$D_{mn} = -\log\left(dis(b_m^{pat}, b_n^{lay})\right) \quad (8)$$

Then, the output of this module is updated through a residual connection with the original patch features to preserve their semantic information.

$$x^{pat+} = x^{pat} + softmax\,(A + IoRs + D)\,V \quad (9)$$

4.3 Triplet Contrastive Loss Module

Following the approach of Zhang et al. (Zhang et al., 2023) in the context of Named Entity Recognition task, we design a Triplet Contrastive Loss to explicitly enhance entity-level consistency between generated captions and reference summaries. The proposed loss function jointly optimizes three core relationships that are crucial for accurate entity generation:

- Word–Word (Word–Word): encourages semantic similarity between predicted entities and their ground-truth counterparts.
- Type–Type (TagType-TagType): enforces consistency between entity types.
- Word–Type (Word-TagType): strengthens the alignment between an entity's semantic representation and its corresponding entity type.

We employ PhoNLP (Nguyen et al., 2021) in combination with VnCoreNLP (Vu et al., 2018) to extract named entities and their corresponding entity types from both the predicted caption and the ground-truth caption. We first concatenate the list of entities (Word) and entity types (TagType) using whitespace separation. Leveraging the representational capacity of BARTpho (Tran et al., 2022), the *embed_tokens* layer of the BARTpho Encoder is used to encode entity sequence and entity-type sequence extracted from the ground-truth caption into dense vector representation. For the predicted caption, entity sequence embedding is obtained

from the *last_hidden_state* of the BARTpho Decoder, while entity-type sequence is encoded using the BARTpho Encoder.

$$\text{Wordseq}^{\text{e}}_{\text{gt}} = \text{BARTpho}^{\text{E}}(\text{Word}_{\text{gt}}) \quad (10)$$

$$\text{Tagtype}^{\text{e}}_{\text{gt}} = \text{BARTpho}^{\text{E}}(\text{Tagtype}_{\text{gt}}) \quad (11)$$

$$\text{Word}^{\text{e}}_{\text{pred}} = \text{BARTpho}^{\text{Decoder-ht}}(\text{Word}_{\text{pred}}) \quad (12)$$

$$\text{Tagtype}^{\text{e}}_{\text{pred}} = \text{BARTpho}^{\text{E}}(\text{Tagtype}_{\text{pred}}) \quad (13)$$

Due to the syllable-based nature of Vietnamese, named entities and numerical expressions often consist of multiple syllables. As a result, single token may fail to capture complete semantic information of word entity. To address this issue, we apply a 1D convolutional layer with syllable window size k over the sequence of feature vectors, allowing each token representation to incorporate contextual information from neighboring syllables.

$$*^{e+}_{\text{gt}} = \text{Conv1d}(*^{e}_{\text{gt}}) \quad (14)$$

$$*^{e+}_{\text{pred}} = \text{Conv1d}(*^{e}_{\text{pred}}) \quad (15)$$

where:

$$*^{\text{e}}_{\text{gt}} = \{\text{Word}^{\text{e}}_{\text{gt}}, \text{TagType}^{\text{e}}_{\text{gt}}\} \quad (16)$$

$$*^{\text{e}}_{\text{pred}} = \{\text{Word}^{\text{e}}_{\text{pred}}, \text{TagType}^{\text{e}}_{\text{pred}}\} \quad (17)$$

To ensure not only fluent generation but also strict entity ordering and correctness, for each entity position i in the ground-truth caption, the model is trained to maximize similarity with a group of syllables/tokens within the range $[i-k, i+k]$ in the predicted caption.

Based on the three defined relationships, we formulate positive contrastive losses as follows:

$$\mathcal{L}_{\text{W-W}} = -\frac{1}{|P_i|}\sum_{i=1}^{N}\log\left(\frac{\sum_{j\in P_i}\exp\left(\text{sim}(w_i^{GT}, w_j^{P})\right)}{\sum_{k\in A_i}\exp\left(\text{sim}(w_i^{GT}, w_k^{P})\right)}\right)$$

$$\mathcal{L}_{\text{T-T}} = -\frac{1}{|P_i|}\sum_{i=1}^{N}\log\left(\frac{\sum_{j\in P_i}\exp\left(\text{sim}(t_i^{GT}, t_j^{P})\right)}{\sum_{k\in A_i}\exp\left(\text{sim}(t_i^{GT}, t_k^{P})\right)}\right)$$

$$\mathcal{L}_{\text{T-W}} = -\frac{1}{|P_i|}\sum_{i=1}^{N}\log\left(\frac{\sum_{j\in P_i}\exp\left(\text{sim}(t_i^{GT}, w_j^{P})\right)}{\sum_{k\in A_i}\exp\left(\text{sim}(t_i^{GT}, w_k^{P})\right)}\right)$$

where

- $\mathcal{L}_{\text{W-W}}$ measures semantic similarity between entities in the predicted caption and the reference summary.
- $\mathcal{L}_{\text{T-T}}$ enforces consistency between entity types.
- $\mathcal{L}_{\text{T-W}}$ aligns entity semantics with their corresponding types.

and

- P_i denotes the set of positive positions
- A_i includes both positive and negative positions.

The overall contrastive loss is computed as:

$$\mathcal{L}_{\text{Contrastive}} = \beta\mathcal{L}_{\text{Word-Word}} + \gamma\mathcal{L}_{\text{Type-Type}} + \delta\mathcal{L}_{\text{Type-Word}} \quad (18)$$

The final training objective combines contrastive learning with the standard cross-entropy loss:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{CE}} + \alpha\mathcal{L}_{\text{Contrastive}} \quad (19)$$

where α, β, γ, and δ are learnable weighting parameters. In particular, α controls the influence of the contrastive objective, preventing the model from over-optimizing entity generation at the expense of fluency and naturalness in the generated captions.

4.4 Multimodal backbone

Following previous works, we adopt a BARTpho (Tran et al., 2022) to encode all the obtained embeddings. S_{source} is embedded by $\text{BARTpho}^{\text{E}}$ layer. The input to the BARTpho backbone is defined as x^f by concatenating x^{pat+} and x^{source}:

$$x^f = [x^{pat+}; x^{source}] \quad (20)$$

4.5 Experiments

4.6 Baselines and Metrics

We compare the performance of multiple representative approaches on the ViInfoSum dataset. **LSTM-R** (Wang et al., 2021b), **M4C-Captioner** (Sidorov et al., 2020), **DEVICE** (Xu et al., 2025), **Anchor Captioner** (Xu et al., 2021), and **LCM-Captioner** (Wang et al., 2023) are state-of-the-art models for OCR-based image captioning, with PhoBERT (Nguyen and Nguyen, 2020) as their backbone for Vietnamese instead of the BERT base (Devlin et al., 2019) for English. **ViT5** (Phan et al.,

#	Method	B@4	M	R	C
1	Anchor Captioner(Xu et al., 2021)	8.41	18.43	22.13	51.52
2	LCM Captioner(Wang et al., 2023)	3.75	18.70	17.28	15.49
3	M4C Captioner(Sidorov et al., 2020)	3.94	19.18	18.19	16.44
4	DEVICE(Xu et al., 2025)	9.49	17.88	23.70	43.45
5	LSTM-R (Wang et al., 2021b)	7.89	16.86	18.91	9.76
6	BARTPho(Tran et al., 2022)	**35.76**	31.30	58.23	286.94
7	ViT5(Phan et al., 2022)	22.49	22.77	37.94	127.99
8	mT5(Xue et al., 2021)	13.26	19.13	40.33	91.72
9	**ViInfoSummer (ours)**	35.42	**31.48**	**59.26**	**306.34**

Table 1: Experimental results of baselines and our proposed ViInfoSummer. The bolded numbers indicated the best scores. *B@4* stands for *BLEU@4*, *M* stands for *METEOR*, *R* stands for *ROUGE*, and *C* stands for *CIDEr*

.

2022), **mT5** (Xue et al., 2021), and **BARTPho** (Tran et al., 2022) serve as strong Vietnamese text summarization backbones.

We assess performance using BLEU@4 (Papineni et al., 2002), METEOR (Lavie and Denkowski, 2009), ROUGE-L (Auriemma Citarella et al.), and CIDEr (Vedantam et al., 2015). While the first two measure word-level overlap and ROUGE-L evaluates semantic alignment, our primary benchmark is CIDEr, as it is the most closely aligned with human intuition and emphasizes the unique, informative tokens that define high-quality outputs.

4.7 Configuration

Patch features are extracted using LayoutLMv3 (Huang et al., 2022), resulting in a total of 196 visual patches per image. Each image is constrained to a maximum of 15 layout regions. The maximum encoder and decoder lengths are set to 128 and 50, respectively. The hidden size is fixed at 1024 to match the hidden dimensionality of BARTPho. The weighting coefficients β, γ, and δ are all set to 0.33, while α is set to 0.5. The syllable window size k is fixed to 3. All models are trained using a batch size of 64 on a single NVIDIA RTX 3090 GPU with 24GB VRAM, employing early stopping. We adopt the AdamW optimizer with an initial learning rate of 1×10^{-4}, apply linear warm-up during the first 10 epochs, and decay the learning rate by a factor of 0.3 every 7 epochs.

5 Experimental Results

Table 1 presents the performance of baseline methods and our proposed *ViInfoSummer* across multiple automatic evaluation metrics.

The first group, consisting of OCR-based image captioning models (Anchor Captioner, LCM Captioner, M4C Captioner, DEVICE, and LSTM-R), achieves relatively low performance on all metrics, especially CIDEr, indicating limited capability to generate semantically rich infographic summaries. Among them, DEVICE performs best, but all OCR-based methods remain substantially inferior to text-based approaches, suggesting that OCR token grounding alone is insufficient for high-level semantic understanding.

The second group includes text summarization models (BARTPho, ViT5, and mT5), which consistently outperform OCR-based methods, demonstrating the importance of strong language modeling for infographic summarization. BARTPho achieves the strongest baseline performance, benefiting from Vietnamese-specific pretraining. However, a clear performance gap remains between these generic summarization models and our proposed approach.

ViInfoSummer achieves the best overall performance, obtaining state-of-the-art results on METEOR, ROUGE-L, and CIDEr. Although its BLEU@4 score is comparable to BARTPho, its substantially higher CIDEr and improvements on recall-oriented metrics demonstrate superior semantic coverage and content fidelity. These results confirm the effectiveness of *ViInfoSummer* in integrating OCR-derived content with high-level semantic reasoning, making it more suitable for Vietnamese infographic summarization than both OCR-based captioning and generic text summarization models.

#	Method	Acc	P	R	F1
1	Anchor Captioner(Xu et al., 2021)	11.36	14.76	13.17	13.00
2	LCM Captioner(Wang et al., 2023)	11.13	16.47	14.10	13.67
3	M4C Captioner(Sidorov et al., 2020)	13.48	17.66	15.43	14.65
4	DEVICE(Xu et al., 2025)	11.18	15.42	13.92	13.46
5	LSTM-R(Wang et al., 2021b)	9.96	21.29	14.18	15.20
6	BARTPho(Tran et al., 2022)	46.19	55.25	**46.64**	48.14
7	ViT5(Phan et al., 2022)	26.51	37.88	29.55	30.56
8	mT5(Xue et al., 2021)	28.01	41.95	28.77	31.55
9	**ViInfoSummer (ours)**	**46.43**	**56.03**	46.56	**48.48**

Table 2: Performance of baselines and our proposed ViInfoSummer on giving correct entities in the summarization of the infographics. Precision (P), Recall (R), and F1 are reported in macro average. *Acc* stands for *Accuracy*.

6 Analysis

Table 2 evaluates the ability of different methods to generate correct named entities in infographic summarization. OCR-based captioning models consistently perform poorly across all metrics, with F1 scores remaining below 16%. Despite direct access to OCR tokens, these models rely primarily on local token-level copying and lack mechanisms for global semantic filtering, making them vulnerable to selecting irrelevant or noisy entities. This limitation is particularly problematic in infographics, where dense textual content often contains multiple entities with varying degrees of relevance.

Text-only summarization models significantly improve entity correctness, indicating that strong language modeling and discourse-level reasoning are essential for accurate entity selection. Among them, BARTPho achieves the strongest performance, reflecting the advantage of Vietnamese-specific pretraining. However, these models operate on flattened textual inputs and therefore lack explicit alignment between textual entities and their visual or spatial context. As a result, they may still include entities that are linguistically plausible but weakly grounded in the underlying infographic structure.

Our proposed *ViInfoSummer* achieves the best overall performance by explicitly addressing these limitations through its architectural design. The Triplet Contrastive Loss encourages the model to learn discriminative semantic representations by pulling summaries closer to relevant content while pushing them away from semantically irrelevant or noisy text, thereby improving entity-level precision. In parallel, the Spatial Overlap Attention module explicitly models the spatial correspondence between OCR text and visual regions, enabling the model to prioritize entities that are both textually salient and spatially grounded. The resulting improvements in Precision and F1 demonstrate that these two components work synergistically: contrastive learning enhances global semantic consistency, while spatial overlap attention refines entity grounding. Together, they enable *ViInfoSummer* to more reliably select and integrate correct entities, leading to more faithful and informative infographic summaries.

7 Conclusion

This paper addresses the task of multimodal summarization of Vietnamese infographics, a challenging problem arising from complex layouts and dense textual content in modern journalism. We constructed a high-quality dataset of 17,840 image–summary pairs collected from Vietnam News Agency infographics, representing the first Vietnamese benchmark dedicated to abstractive summarization under multimodal settings. Comprehensive statistical analyses on domain distribution, text length, lexical diversity, and OCR–caption correlation reveal a dataset rich in named entities and numerical information.

Moreover, we propose a novel multimodal summarization method for Vietnamese ViInfoSummer that explicitly models spatial layout relationships and incorporates a triplet contrastive loss to enhance the factual accuracy of entities and numbers in generated summaries. Experimental results demonstrate that the proposed design effectively mitigates hallucination and improves the reliability of multimodal summarization, providing a strong foundation for future research on text-rich visual understanding in Vietnamese.

Acknowledgment

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number NCM2025-26-02.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Alessia Auriemma Citarella, Marcello Barbella, Madalina G Ciobanu, Fabiola De Marco, Luigi Di Biasi, and Genoveffa Tortora. Rouge metric evaluation for text summarization techniques. *Available at SSRN 4753220*.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025. Paddleocr 3.0 technical report. *Preprint*, arXiv:2507.05595.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chengyang Fang, Jiangnan Li, Liang Li, Can Ma, and Dayong Hu. 2023. Separate and locate: Rethink the text in text-based visual question answering. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4378–4388.

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 (CVPR)*, pages 14867–14878.

Ronghang Hu, Amanpreet Singh, Anna Rohrbach, Manyipuiedu Anyigbo, and Devi Parikh. 2020. Iterative answer prediction with truncated decoding for text visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM international conference on multimedia*, pages 4083–4091.

Alon Lavie and Michael J Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine translation*, 23(2):105–115.

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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Dat Quoc Nguyen and Anh-Tuan Nguyen. 2020. Phobert: Pre-trained language models for vietnamese. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 1037–1042.

Dat Quoc Nguyen et al. 2021. Phonlp: A joint multi-task learning model for vietnamese part-of-speech tagging, named entity recognition and dependency parsing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 1–7.

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.

Long Phan, Hieu Tran, Hieu Nguyen, and Trieu H. Trinh. 2022. ViT5: Pretrained text-to-text transformer for Vietnamese language generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 136–142, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *European conference on computer vision*, pages 742–758. Springer.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Look, read, reason & answer: Learning to answer questions in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Nguyen Luong Tran, Duong Le, and Dat Quoc Nguyen. 2022. BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese. In *Interspeech 2022*, pages 1751–1755.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. *Advances in neural information processing systems*, 28.

Thanh Vu, Dat Quoc Nguyen, Mark Dras, Mark Johnson, et al. 2018. Vncorenlp: A vietnamese natural language processing toolkit. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: demonstrations*, pages 56–60.

Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. 2021a. Screen2words: Automatic mobile ui summarization with multimodal learning. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pages 498–510.

Jing Wang, Jinhui Tang, Mingkun Yang, Xiang Bai, and Jiebo Luo. 2021b. Improving ocr-based image captioning by incorporating geometrical relationship. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1306–1315.

Qi Wang, Hongyu Deng, Xue Wu, Zhenguo Yang, Yun Liu, Yazhou Wang, and Gefei Hao. 2023. Lcm-captioner: A lightweight text-based image captioning method with collaborative mechanism between vision and text. *Neural Networks*, 162:318–329.

Dongsheng Xu, Qingbao Huang, Xingmao Zhang, Haonan Cheng, Feng Shuang, and Yi Cai. 2025. Device: Depth and visual concepts aware transformer for ocr-based image captioning. *Pattern Recognition*, 164:111522.

Guanghui Xu, Shuaicheng Niu, Mingkui Tan, Yucheng Luo, Qing Du, and Qi Wu. 2021. Towards accurate text-based image captioning with content diversity exploration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12637–12646.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, pages 2048–2057.

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.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2021. Tap: Text-aware pre-training for text-vqa and text-captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15562–15572.

Sheng Zhang, Hao Cheng, Jianfeng Gao, and Hoifung Poon. 2023. Optimizing bi-encoder for named entity recognition via contrastive learning. In *International Conference on Learning Representations (ICLR)*.

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.

Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2021. Graph-based multimodal ranking models for multimodal summarization. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(4).

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

A Ablation Study Components in ViInfoSummer

Table 3: Ablation study on the Spatial Overlap Attention (SOA) module and the Triplet Contrastive Loss (UCL) of the proposed ViInfoSummer.

#	SOA	TCL	BLEU@4	METEOR	ROUGE-L	CIDEr
1	✗	✗	**35.76**	31.30	58.23	286.94
2	✗	✓	35.42	31.38	59.11	305.58
3	✓	✓	35.42	**31.48**	**59.26**	**306.34**

Table 3 presents the ablation study evaluating the contributions of the Spatial Overlap Attention (SOA) module and the Triplet Contrastive Loss in *ViInfoSummer*. Starting from the base model without either component, incorporating the Triplet Contrastive Loss alone leads to consistent improvements across all metrics, most notably on ROUGE-L and CIDEr. This indicates that contrastive supervision effectively enhances semantic alignment between generated summaries and reference captions by encouraging better discrimination of salient information. The gains observed with contrastive learning suggest that global semantic consistency

plays a critical role in producing high-quality infographic summaries.

Adding the Spatial Overlap Attention module on top of the contrastive objective further improves performance, yielding the best results on METEOR, ROUGE-L, and CIDEr. This demonstrates that explicitly modeling spatial overlap between visual regions and textual elements helps the model more accurately localize and aggregate relevant information. While BLEU@4 remains relatively stable across configurations, the improvements in recall-oriented and consensus-based metrics indicate stronger content coverage and semantic fidelity. Overall, the ablation results confirm that the two components are complementary: contrastive learning enhances global semantic representation, while Spatial Overlap Attention refines fine-grained multimodal alignment, together contributing to the superior performance of *ViInfoSummer*.

B Ablation Study on the weight of Triplet Contrastive Loss α in the Objective Function

#	α	BLEU@4	METEOR	ROUGE-L	CIDEr
1	0.0	**36.05**	**31.56**	58.76	288.81
2	0.3	34.87	31.02	57.49	295.29
3	0.5	35.42	31.48	**59.26**	**306.34**
4	0.7	35.19	31.08	58.09	296.63

Table 4: Ablation study on the factor α of the Triplet Contrastive Loss.

According to Table 4, when $\alpha = 0$, the model is trained solely with cross-entropy loss and achieves strong BLEU@4 and METEOR scores, indicating effective token-level generation. However, its CIDEr score is noticeably lower than configurations that incorporate contrastive learning, suggesting limited ability to capture global semantic consensus with human-written summaries.

Introducing contrastive learning with a moderate weight leads to consistent improvements in semantic-oriented metrics. In particular, setting $\alpha = 0.5$ yields the best overall performance, achieving the highest ROUGE-L and CIDEr scores while maintaining competitive BLEU@4 and METEOR values. This demonstrates that the Triplet Contrastive Loss effectively enhances high-level semantic alignment and content selection when appropriately balanced with the generation objective.

By contrast, assigning too small a weight to the contrastive term ($\alpha = 0.3$) provides insufficient semantic regularization, while an overly large weight ($\alpha = 0.7$) slightly degrades performance across metrics, likely due to overemphasizing representation separation at the expense of fluent sequence generation.

Overall, these results indicate that contrastive learning acts as a complementary supervisory signal, and that $\alpha = 0.5$ offers an optimal trade-off between local generation fidelity and global semantic consistency in *ViInfoSummer*.

C Effect of OCR on the Performance of ViInfoSummer

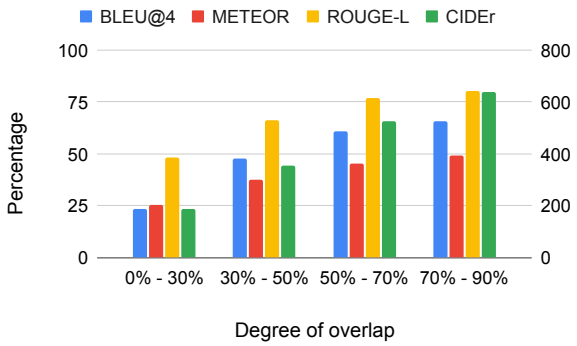


Figure 6: Results of the proposed ViInfoSummer according to the percentage of detected OCR tokens in the ground truth summarization.

Figure 6 shows that the performance of ViInfoSummer consistently improves as the overlap between OCR-extracted tokens and the ground-truth summaries increases. All metrics (BLEU@4, METEOR, ROUGE-L, and CIDEr) rise monotonically from the low-overlap range (0–30%) to the high-overlap range (70–90%). This trend indicates that ViInfoSummer can effectively exploit OCR information when textual cues in the infographic are relevant to the target summary, as well as demonstrate the important of scene texts for multimodal summarization on the ViInfoSum dataset.

D Quantitative Analysis

The quantitative results in Figure 7 reveal a clear performance gap between *BARTPho* and the proposed *ViInfoSummer*, underscoring the limitations of applying generic text summarization models to infographic summarization. Despite its strong Vietnamese pretraining, *BARTPho* consistently underperforms on all metrics, particularly those sensitive to factual overlap and content fidelity. This

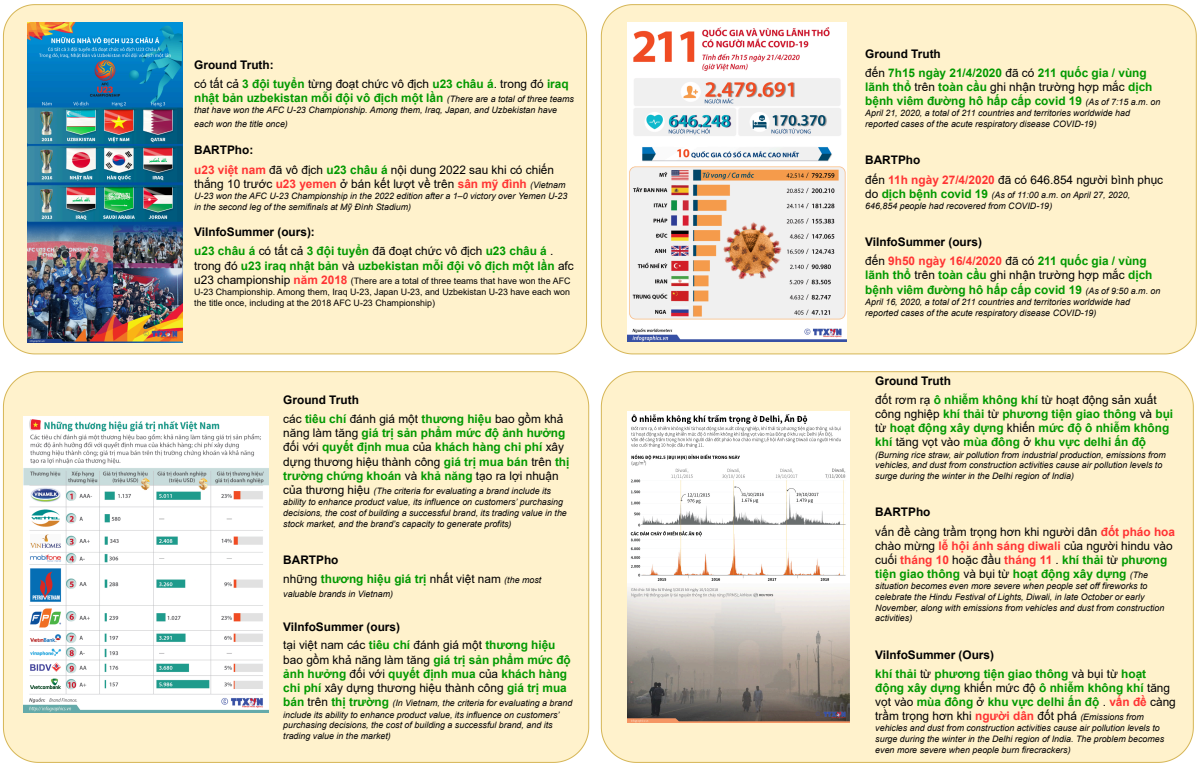


Figure 7: Quatitative results of the BARTPho and our proposed ViInfoSummer. Green texts indicate the correct entities while red texts denote the incorrect entities in the generated summarization.

behavior indicates that fluency-driven generation alone is insufficient in this setting, as *BARTPho* frequently produces plausible yet factually incorrect or incomplete summaries due to its lack of explicit grounding in visually embedded textual information. The resulting degradation in quantitative scores highlights a fundamental mismatch between text-only summarization objectives and the information-centric nature of infographic data.

In contrast, *ViInfoSummer* demonstrates substantially stronger quantitative performance by explicitly addressing the mismatch of entities through contrastive learning and spatial information from the visual context. By aligning OCR-derived content with summary generation and enforcing semantic discrimination during training, *ViInfoSummer* significantly improves content selection, factual precision, and global semantic consistency. The consistent gains across all evaluation metrics confirm that infographic summarization requires more than linguistic competence; it demands explicit mechanisms for information extraction, alignment, and aggregation. These results establish *ViInfoSummer* as a fundamentally more suitable paradigm for Vietnamese infographic summarization, rather than a marginal extension of existing text summarization models.