BV-FRD: A Multimodal Vietnamese-English Food Review Video Description Generation

Bao Pham-Thai, Vy Do Le Khanh, Huy Quoc To

Proceedings of the 39th Pacific Asia Conference on
Language, Information and Computation (PACLIC 39)

Emmanuele Chersoni, Jong-Bok Kim (eds.)

2025

BV-FRD: A Multimodal Vietnamese-English Food Review Video Description Generation

Bao Pham-Thai[1,2], Vy Do Le Khanh[1,2], Huy Quoc To[1,2]
[1]University of Information Technology, Ho Chi Minh city, Vietnam
[2]Vietnam National University, Ho Chi Minh city, Vietnam
{21520156, 23521826}@gm.uit.edu.vn, huytq@uit.edu.vn

Abstract

The short video summarization provides brief descriptions of the main content, enabling viewers to quickly understand the key information. Although the field has seen significant progress, there remains a shortage of datasets for food reviews, particularly in Vietnamese. In this paper, we introduce a novel multimodal Vietnamese-English dataset focused on Vietnamese food review videos called **BV-FRD**. Our dataset includes a wide range of food, prices and restaurant locations. Each video includes processed scripts and annotated Vietnamese-English descriptions, generated through a multi-stage pipeline using several LLMs with human collaboration. Baseline experiments show moderate performance, indicating that the dataset is challenging and has strong potential for practical applications. In our experiments, DeepSeek outperforms other models in Vietnamese and English across three of four evaluation metrics. In Vietnamese, Phi-4 achieves the highest BERTS score, with a value of 0.64 precision. In English, DeepSeek reaches the highest consistency in Uni-Eval, with a value of 0.78. Through our analysis and baseline experiments, we demonstrate that our dataset is valuable and challenging for multimodal food review description generation task. Our dataset is available through this link[1].

1 Introduction

Every day, millions engage with short videos related to food, shaping food trends, domestic tourism, and global perceptions of Vietnamese cuisine (Truong and Kim, 2023). These videos on social networks attract the audience through their brevity, practicality, and high shareability (Violot et al., 2024; Zannettou et al., 2024). In addition, it can affect consumers' choices, identity construction, and cultural preservation.

Most of this content is open source, providing multi-modal datasets that combine speech, facial expressions, visuals, and sounds, offering a clearer view of real-world communication than text-only data (Baltrušaitis et al., 2018). Open-source data promotes transparency, reproducibility, and supports research in artificial intelligence and human–computer interaction (Von Krogh et al., 2003; Willmes et al., 2014).

In recent years, Vietnamese natural language processing (NLP) has advanced significantly, with high-quality studies on word segmentation (Hai et al., 2025), document summarization (Le and Le, 2025), and social media text processing (Nguyen et al., 2023). Pre-trained models like PhoBERT (Nguyen and Nguyen, 2020) have boosted performance across downstream tasks. Research on sentiment analysis and question-answering systems (Lê et al., 2020) further enriches the NLP landscape. Overall, these developments lay the foundation for advanced Vietnamese language models capable of contextual understanding, supporting applications from automated customer support to content creation.

Although a niche domain, food reviews strongly influence consumers, promote local cuisine, and support cultural exchange (Rini et al., 2024). Research on short Vietnamese videos is limited, and open-source multimodal datasets combining images, audio, and text are lacking, which are crucial for conversational AI and multimodal understanding (Baltrušaitis et al., 2018).

We present the Vietnamese–English bilingual dataset **BV-FRD** for short food-related videos, containing transcripts, manually written summaries, and emotion labels to enrich Vietnamese language resources and support research on dialogue summarization, multimodal sentiment analysis, and contextual understanding. The main task is to summarize food review videos on YouTube Shorts into short, easy-to-understand descriptions. The final re-

[1]https://anonymous.4open.science/r/BV-FRD

sult is a dataset of 2,020 videos, processed through a pipeline using several LLMs with manual verification at each step. The dataset is diverse in dishes, pricing, and restaurant locations, with additional information such as video duration, view counts, and temporal distribution, ensuring representativeness and usefulness for various downstream tasks.

Our dataset makes two main contributions:

- In this paper, we introduce the BV-FRD dataset. We leverage multiple LLMs in the processing stage to enhance data generation and employ a human-in-the-loop process to ensure high-quality outputs. As a result, the dataset contains 2,020 samples with a total duration of 31.75 hours. Our dataset provides a diverse and reliable resource for generating Vietnamese-English text summaries of food review videos.

- We evaluate our dataset with four base models on four metrics in both English and Vietnamese. For both languages, DeepSeek achieved the highest performance on three of the four metrics. However, the generated descriptions are only relatively similar to ground truth. The performance of these baseline models indicates that our dataset is challenging and has potential for further research in summarizing food review videos into bilingual description.

The presentation of our paper is organized as follows. **Section 2** presents related works on open-source releases concerning videos. **Section 3** introduces the video processing pipeline for generating bilingual Vietnamese and English descriptions, with steps involving both human annotators and LLMs. **Section 4** analyzes our dataset, including statistical information and diversity measures. **Section 5** presents the results of applying base models to tasks based on our dataset. **Section 6** concludes the work conducted and outlines directions for future development.

2 Related Work

In the multimodal era, video has become a dominant medium for information dissemination, driving the growing demand for efficient video comprehension and summarization (Apostolidis et al., 2021; Otani et al., 2022). Existing datasets support diverse tasks such as video captioning (Wang et al., 2019) ,summarization (Qiu et al., 2024), and meeting conversation analysis (Carletta et al., 2005). Notably, MSR-VTT (Xu et al., 2016) offers daily-life videos with short descriptions, VATEX (Wang et al., 2019) adds bilingual captions, and MMSum (Qiu et al., 2024) incorporates dialogue and metadata. While effective for short, visually simple clips, these datasets lack coverage of narrative-rich content such as vlogs or culinary reviews. To fill this gap, we construct a dataset with varied description lengths that more accurately captures video semantics, enabling adaptation to tasks from fine-grained understanding to concise summarization and enhancing model robustness.

In the food review domain, the NLP community in English benefits from large-scale datasets like Amazon Fine Food Reviews (McAuley et al., 2015) and Yelp Reviews (Ganu et al., 2009), supporting sentiment analysis, opinion summarization, and multimodal tasks. Vietnamese research has also progressed, with datasets from Foody and Lozi used for sentiment analysis and classification. (Nguyen et al., 2021) collected over 236k reviews from 2011–2020, achieving 91.5% sentiment classification accuracy. ViMRHP (Nguyen et al., 2025) introduced a multimodal dataset for helpfulness assessment. However, most remain monolingual, limiting cross-cultural accessibility. Our Vietnamese–English bilingual dataset fills the resource gap while enhancing the global presence of Vietnamese cuisine. The aligned bilingual data facilitates the development of multilingual models and advances NLP research for Vietnamese—a low-resource language. Additionally, it supports the promotion of Vietnamese culinary culture to a wider international audience, contributing to the preservation and development of local heritage.

NLP has been applied across domains with datasets such as FFVD for e-commerce product review videos (Zhang et al., 2020), Video Story for social media narrative interaction and emotion analysis (Gella et al., 2018). While these resources advance research, their manual collection and annotation incur high costs and limit scalability (Yuan et al., 2025). Similar challenges appear in datasets like VideoCC (Nagrani et al., 2022), where LLMs generate detailed captions to reduce effort and cost, but fully automated approaches risk semantic gaps and inconsistencies (Liu and Wan, 2023). To address these issues, we adopt a semi-automated pipeline using LLMs for preprocessing and annotation, followed by rigorous human verification to

Table 1: Summary of Related Work Datasets

Dataset	Domain	Year	Source	Language	Type	Annotation	Human Verify	LLM	Location Variety
MSR-VTT (Xu et al., 2016)	Multi–category	2016	Youtube + AMT	EN	Caption	Crowdsourcing (AMT)	✓	✗	✗
ActivityNet Caption (Krishna et al., 2017)	Human Activity	2017	ActivityNet (YouTube)	EN	Description (dense)	Crowdsourcing (AMT)	✗	✗	✗
YouCook II (Zhou et al., 2018)	Cooking	2018	Youtube	EN	Caption	Human	✓	✗	✗
VATEX (Wang et al., 2019)	Multi–category	2019	Kinetics-600 dataset (YouTube)	EN + ZH	Caption	Crowdsourcing (AMT)	✓	✗	✗
VideoCC (Nagrani et al., 2022)	Multi–category	2022	YouTube	EN	Caption (auto)	LLM (auto)	✗	✓	✗
TACoS-MLevel (Rohrbach et al., 2014)	Cooking	2018	AMT + TACoS	EN	Description	Human	✓	✗	✗
VideoStory (Gella et al., 2018)	Social Media	2018	Social media platform	EN	Multi-sentence description	Human	✗	✗	✗
FFVD (Zhang et al., 2020)	E-Commerce	2020	Mobile Taobao	EN	Caption	Human	✓	✗	✗
MMSum (Qiu et al., 2024)	MultiModel Summarization	2024	Youtube	EN	Caption	Human	✓	✗	✗
Ours	**Review Food**	**2025**	**Youtube**	**VI + EN**	**Description**	**LLM + Human**	✓	✓	✓

ensure accuracy, reliability, and scalability.

Prior studies show that product attributes—such as dish name, price, and restaurant location—play a decisive role in consumer choices (FUENTES). In culinary video datasets, YouCookII (Zhou et al., 2018) captured diverse recipes from multiple continents, while TACoS Multi-Level (Rohrbach et al., 2014) focused on detailed cooking steps and ingredients. Beyond content, video popularity metrics (view count, watch time, comments) strongly influence engagement and trust (Park et al., 2016; Liu et al., 2025). Building on these insights, we define six criteria to assess diversity in food review datasets, aiming to give users a comprehensive overview of available resources.

BV-FRD has been developed as a Vietnamese–English bilingual resource through a semi-automated process designed to improve the efficiency of data collection and processing, as shown in **Table 1**. The dataset aims to address the resource scarcity for Vietnamese, enhance data quality and scalability, and promote Vietnamese culinary culture internationally while supporting research in natural language processing and artificial intelligence.

3 BV-FRD Dataset Creation

Firstly, the data is collected from YouTube short videos, specifically focusing on Vietnamese-language food review content. We then extract video transcripts and apply the GPT-4 model to refine them into a version focusing solely on food review information. Next, our annotators verify and edit the output. The Gemini model is subsequently employed to summarize the refined transcripts into descriptions, followed by another round of human verification. For English translation, the VinAI model is used, with additional verification by GPT-4. Human checks are incorporated at all stages to maintain factual accuracy and contextual consistency, as detailed in **Appendix A.1**. The complete workflow is illustrated in **Figure 1**. Prompt templates and model configurations for each stage—script refinement, description creation, and translation—are provided in **Appendix A.2**.

3.1 Short Video Collection

The videos are collected exclusively from YouTube in the form of short videos. The script information is obtained from transcripts publicly provided by the video uploaders, and we select channels that already include such scripts.

We only consider channels and videos that pro-

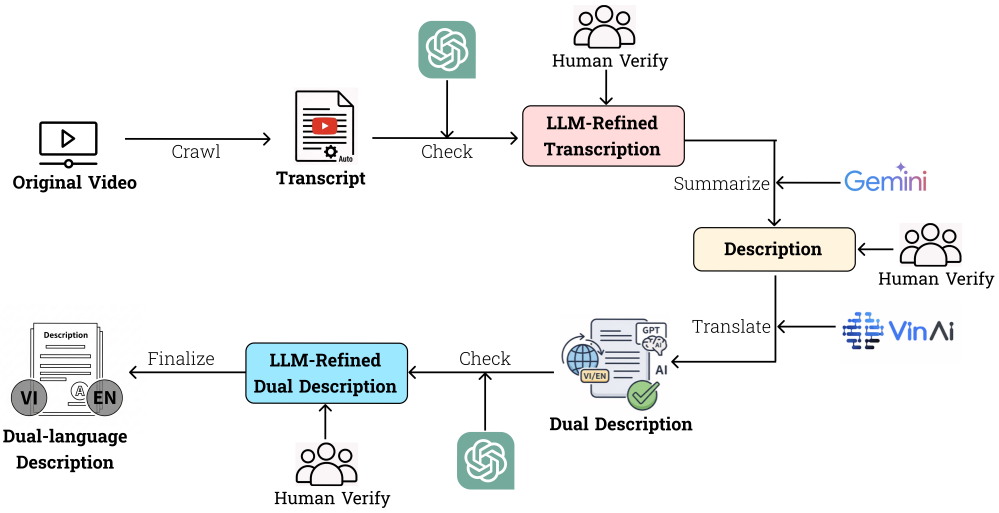

Figure 1: Video-to-Text Dataset Construction Pipeline

Figure 2: Visual evidence of contextual elements absent from the textual description

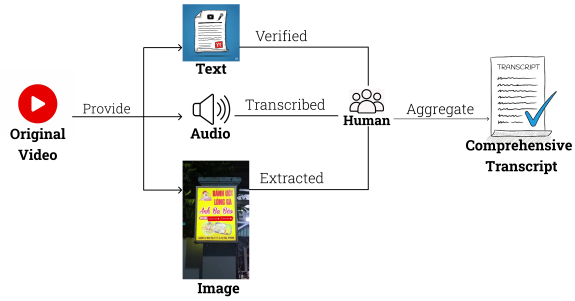

Figure 3: Comprehensive Transcript Generation Workflow

duce content related to food reviews. Furthermore, each collected short video is manually filtered to ensure that its content is entirely relevant to food review topics.

3.2 Script Processing and Description Generation

Based on prior work demonstrating GPT-4's strong capability in processing multimodal inputs, including text, images, and videos (Alam et al., 2024), we employ GPT-4 to understand the context and main content of videos, allowing accurate and efficient identification of food-related information. Therefore, with the collected scripts, we use GPT-4 to extract the core content specifically related to food reviews. Then, human annotators verify and refine the extracted information to ensure accuracy. Although the initial data is in textual form, the processing pipeline is further extended to incorporate both image and audio information, thereby ensuring a more comprehensive representation of the content. As shown in certain parts of the content, as illustrated in **Figure. 2**, may also appear in image form, making it necessary to process both images and audio to supplement the script with further details.

The complete workflow for processing these supplementary modalities along with the script is illustrated in **Figure. 3**. In this process, we use the GPT-4 model as an LLM to understand both the input and the video context, enabling the extraction of the required review-focused content. Human participation is essential to enrich the extracted information with visual and auditory details, correct spelling errors, supplement missing information, and remove irrelevant elements that do not align with the context of food review.

The use of two different LLMs is intentional. GPT-4 is used for extraction, while Gemini is used for summarization. Prior research has shown that users may over-rely on a single LLM, leading to uncritical acceptance of confidently stated but in-

correct outputs (Bo et al., 2025). Employing two LLMs helps avoid over-reliance on a single model and mitigates model-specific biases and hallucinations. Additionally, the two models complement each other's strengths: GPT-4 excels at precise content extraction, while Gemini produces fluent and coherent summaries. The human-in-the-loop component is essential for ensuring data accuracy, contextual alignment, and editorial quality. The dual-LLM plus human strategy enhances diversity, reliability, and prevents errors. Prior studies also show that LLMs can hallucinate, producing fluent but incorrect content, so human verification ensures dataset authenticity (Huang et al., 2025). Based on prior reports highlighting Gemini's strong multimodal understanding capabilities, including long-context processing of up to three hours of video content ((Comanici et al., 2025); (Akter et al., 2023)), We employ Gemini to capture the overall context and key points of a document, enabling accurate and fluent summaries of script content in the domain of food reviews. Human annotators who are undergraduate students then verify the accuracy, completeness, and contextual relevance of the generated descriptions, adding or adjusting the information when necessary.

3.3 Bilingual Description Generation

The video description is originally in Vietnamese; to broaden accessibility, we generate an English version. The VinAI model, which supports both English and Vietnamese language processing (Tran and Thanh, 2024), is used to produce the English translation, making it a suitable choice for our task. Such characteristics are crucial for food review content, where cultural context and descriptive precision are important. After translation, verification and editing are performed using the GPT-4 model. (Yan et al., 2024) presented promising research on the use of LLMs for translation, including GPT-4, with results demonstrating its superior effectiveness. Therefore, we employ GPT-4 to assess the contextual relevance and fidelity of the translation, using both Vietnamese and English descriptions for direct comparison. Finally, a human annotator reviews the output, making revisions or additions if necessary.

In our paper, we use GPT-4 to extract information and check the translation of descriptions because it handles many tasks well, especially with multiple languages (Blake, 2025). However, we also use other models and have humans review the results to ensure accuracy. This combination ensures that the dataset is not overly dependent on the information processed solely by GPT-4, maintaining robustness and diversity in the data processing pipeline.

4 BV-FRD Analysis

This section presents a detailed analysis of the dataset's characteristics, including statistical summaries and diversity assessment. Emphasis is placed on ensuring the dataset covers a wide range of scenarios, which is crucial for improving model robustness and enabling reliable performance evaluation.

4.1 Dataset Statistics

The detailed characteristics of the proposed dataset are presented in **Table 2**. It contains a large number of videos, with over seven thousand collected and more than two thousand processed. The total duration reaches 31.75 hours, indicating that the dataset spans a wide range of content lengths. The mean duration per video is 56.59 seconds, which is sufficient to capture the concise style typical of YouTube Shorts. Examples of selected samples are presented in **Appendix A.3**. To ensure quality, the collection process relies on channel information containing videos related to food reviews. We carefully select videos that focus on food review content, while excluding those with little relevance to the domain.

4.2 Diversity Analysis

In the food domain, several related datasets exist (Zhou et al., 2018; Das et al., 2013; Regneri et al., 2013; Rohrbach et al., 2014; Huang et al., 2020). Our dataset targets the food review context with bilingual (Vietnamese–English) descriptions. It is built through a multi-LLM pipeline with human verification, and carefully curated to ensure diversity in video content. As summarized in **Table 3**, it is diverse in both content and language, while remaining independent of specific LLMs or human reviewers.

Table 2: Video Data Statistics

Total Number of Videos Collected	7292
Total Number of Videos Processed	2020
Aggregate Duration (hours)	31.75
Mean Duration per Video (seconds)	56.59

Table 3: Benchmark Datasets in the Food Review Domain

Dataset	Source	Billigual	Human Verified	LLM Used	Food Variety	LLM Usage & Steps
YouCook II (Zhou et al., 2018)	Youtube	✗	✓	✗	✓	0
YouCook (Das et al., 2013)	Youtube	✗	✗	✗	✓	0
TACoS (Regneri et al., 2013)	MPII Cooking Composite Activities	✗	✓	✗	✗	0
TACoS-MLevel (Rohrbach et al., 2014)	AMT + TACoS	✗	✓	✗	✓	0
ViTT (Huang et al., 2020)	YouTube-8M + Ann	✗	✗	✓	not mention	1 & 1
Ours	**Youtube**	✓	✓	✓	✓	**3 & 4**

A more detailed presentation of the statistical results, along with a comprehensive analysis of the dataset's diversity across different quantities, is provided in **Appendix A.6**, where the variations, distributions, and underlying patterns are thoroughly examined to support the findings discussed in the main text.

Video duration affects viewer engagement. As shown in **Figure. 4**, most videos balance information delivery and attention span, with the majority (1,694) lasting –60 seconds. **Figure. 5** shows a diverse distribution of view counts, reflecting real-world user behavior. Most videos have fewer than 10,000 views, whereas 128 exceed 1,000,000, indicating a subset of highly viral content.In **Figure. 6**, most selected videos were posted in recent years, ensuring that the dataset aligns with current social media trends and user interests. Notably, the videos span from 2020 to 2025, with as many as 639 uploaded in 2025. This diversity enhances the reliability and applicability of the research results.

The food characteristics of the dataset are highlighted by the frequent appearance of signature dishes, such as Pho, as shown in **Figure. 7**. This reflects both cultural significance and the realism of actual user culinary experiences. The data provide a broad overview while accurately representing popular dishes in everyday life, enhancing au-

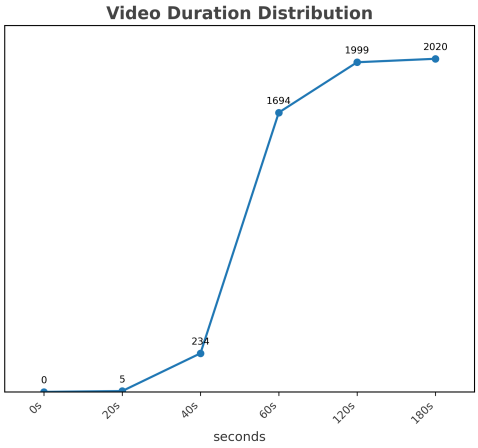

Figure 4: Distribution of video durations

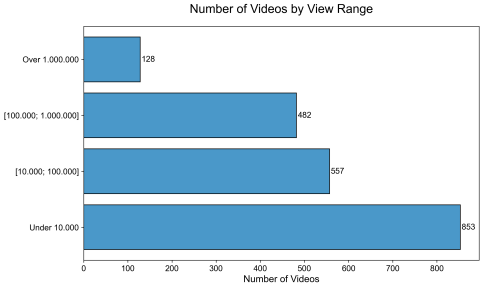

Figure 5: Distribution of video views

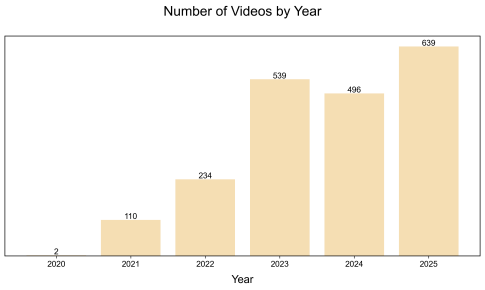

Figure 6: Distribution of release year

Figure 7: Word cloud of food name variety

thenticity and reliability for analysis and model development. Geographically, the cuisine spans multiple regions, with Ho Chi Minh City accounting for a substantial share of 1,723 videos due to its socio-economic prominence, as illustrated in **Figure. 8**. This diversity captures regional differences in culinary styles and user behaviors, improving research comprehensiveness and reducing geographical bias. The dataset reflects diverse price ranges, as illustrated in **Figure. 9**. This distribution supports analyses on price–experience relationships and applications such as food recommendation and market studies.

The analysis and synthesis of these criteria not only enhance the practical value of the dataset, but also provide a clear theoretical framework, serving as a foundation for future academic research in the field of food review. These standards help ensure that data achieve realism, diversity, and contextual appropriateness, helping to increase the reliability and effectiveness of machine learning models, data analysis, as well as applications in the development of recommendation systems or automatic evaluation tools.

5 Experiments

We evaluate our dataset for the food review video description generation task using multiple baseline LLMs and metrics, highlighting its challenge and practical value.

5.1 Experimental Setup

The processing workflow with GPT-4[2] and Gemini[3] models uses an API Key for connection. We provide the input content along with a prompt to guide the model in producing the desired output.

[2] https://chatgpt.com/
[3] https://gemini.google.com/app

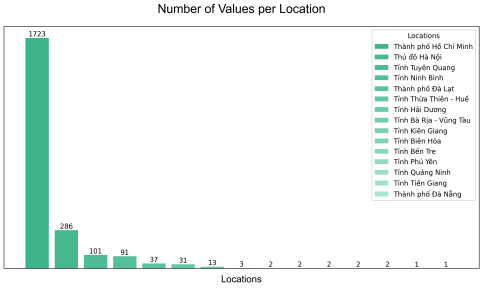

Figure 8: Geographic distribution of restaurants in the dataset

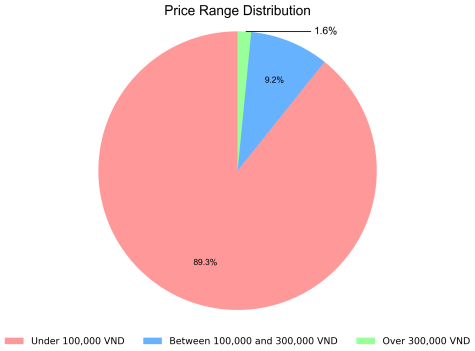

Figure 9: Distribution of food prices in the dataset

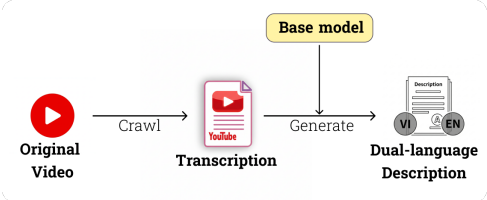

Figure 10: Food review video description generation task on our dataset

Meanwhile, we apply the VinAI model in version 2 with the vi2en[4] mode.

In addition, we employ base LLM models to evaluate our dataset by generating concise Vietnamese descriptions from auto-generated video transcripts and translating them into English, leveraging their strong text processing and contextual understanding. We use four open-source LLMs: Gemma (Team et al., 2024), Qwen (Yang et al., 2025), DeepSeek (DeepSeek-AI et al., 2025), and Phi-4 (Abdin et al., 2024). The versions of each model used in our experiment are listed in **Appendix A.4**. Our experiments cover from medium-size (7B) to large-size (14B) architectures for comparative analysis. Their strong reasoning, text generation, and multilingual abilities make them suitable for generating and translating video descriptions. The experiments are run on a P40 system with an Intel Xeon E5-2680 v3 CPU (20 cores), 48 GB memory, 300 GB storage, and specified network bandwidth.

The bilingual food review description generation task requires a LLM model to use the video transcription content as input, and the expected output is the bilingual Vietnamese and English description, as illustrated in **Figure. 10**.

[4] https://huggingface.co/vinai/vinai-translate-vi2en

5.2 Evaluation Metrics

We evaluated the generated descriptions using BERTScore (Zhang et al., 2019), PhoBERT (Nguyen and Nguyen, 2020), UniEval (Zhong et al., 2022), and ROUGE (Lin, 2004), covering semantic similarity, linguistic quality, lexical overlap, and translation adequacy. BERTScore assesses semantic alignment via Precision, Recall, and F1-score. PhoBERT, specialized for Vietnamese, ensures accurate source-language evaluation. UniEval measures consistency and relevance for contextual coherence. ROUGE-L captures the longest common subsequence (LCS) between the candidate and reference texts. We report F1-score as a balanced metric. Collectively, these metrics provide a comprehensive framework to evaluate bilingual food review descriptions, ensuring accuracy, relevance, and fidelity in both Vietnamese and English.

5.3 Results

The evaluation results for Gemma, Qwen, DeepSeek, and Phi-4 in the generation of Vietnamese descriptions are shown in **Table 4**. DeepSeek achieves the highest Recall (0.63), ROUGE (0.08), and PhoBERT (0.63), while Phi-4 attains the highest Precision (0.64). Gemma and Qwen perform moderately, with Precision and Recall around 0.60 and ROUGE 0.06. Overall, all models show modest performance, highlighting the difficulty of the bilingual food review dataset and its value as a benchmark for future research. Common errors observed across models include incomplete or truncated descriptions, lexical repetition, misuse of domain-specific terms, and occasional semantic drift where generated text deviates from the actual video content. These issues suggest challenges in both content grounding and maintaining linguistic accuracy in Vietnamese.

The performance of Gemma, Qwen, DeepSeek, and Phi-4 on English description generation from video transcripts is summarized in **Table 5**. DeepSeek achieves the highest Relevance (0.78), ROUGE (0.11), and F1-Score (0.65), while Qwen leads in Consistency (0.55) and ranks second overall. Gemma shows moderate performance, and Phi-4 records the lowest scores, with the largest gap in Relevance (0.60) between DeepSeek and Phi-4. Common errors include omission of key contextual details, overly generic or repetitive phrasing, and inconsistencies with the transcript, reflecting challenges in bilingual video description. Limitations such as insufficient data diversity, weak generalization to unseen content, and potential overfitting further affect robustness, underscoring the dataset's value as a benchmark for future research.

Some examples of model outputs for our experiments are presented in **Appendix A.5**.

Table 4: Evaluation Metric For Vietnamese Text. BS-P is BERTScore-Precision, BS-R is BERTScore-Recall, ROUGE is F1-score of ROUGE-L.

Model	BS-P	BS-R	ROUGE	PhoBERT
Gemma	0.62	0.60	0.06	0.50
Qwen	0.61	0.59	0.06	0.60
DeepSeek	0.62	**0.63**	**0.08**	**0.63**
Phi-4	**0.64**	0.51	0.02	0.40

Table 5: Evaluation Metric For English Text. Uni-Eval-C is Consistency of Uni-Eval, Uni-Eval-R is Relevancy of Uni-Eval, ROUGE is F1-score of ROUGE-L, BS-F1 is BERTScore-F1-score.

Model	Uni-Eval-C	Uni-Eval-R	ROUGE	BS-F1
Gemma	0.50	0.43	0.10	0.63
Qwen	**0.55**	0.62	0.10	0.62
DeepSeek	0.53	**0.78**	**0.11**	**0.65**
Phi-4	0.44	0.18	0.07	0.61

6 Conclusion

We release a food review dataset, analyzing video quality and diversity in dish, price, and location, focusing on Vietnamese content and English translation. Processed via an LLM–Human pipeline, the dataset ensures high-quality information. Each step—LLM, Human, or both—is documented and applicable to various tasks. The dataset supports generating Vietnamese and English descriptions from transcripts. Base model performance is relatively low, highlighting the dataset's challenge and the need for further research.

In the future, we will continue to develop the dataset with a larger number of videos. We aim to build appropriate processing models and achieve effective results on the dataset we publish.

Limitations

Our work builds a bilingual dataset for food reviews using Vietnamese-language videos as input. The pipeline combines human expertise with LLM capabilities. However, the dataset size is still limited, reducing generalizability. Information extraction from visual and auditory content depends entirely on humans, which may cause errors. There is no

direct user evaluation of the generated descriptions. The pipeline also lacks automated multimodal analysis, which could improve efficiency and scalability in future work.

Ethics Statement

This study uses a dataset of publicly available food review videos collected in compliance with platform terms of service. No personally identifiable information was collected, and sensitive content was removed during preprocessing. The dataset, containing only research-relevant text and metadata, is used solely for non-commercial academic purposes. All processes follow privacy-by-design principles and include safeguards to prevent any potential harm to users or their data.

Acknowledgments

This research is funded by University of Information Technology, Vietnam National University, Ho Chi Minh City, Vietnam.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Syeda Nahida Akter, Zichun Yu, Aashiq Muhamed, Tianyue Ou, Alex Bäuerle, Ángel Alexander Cabrera, Krish Dholakia, Chenyan Xiong, and Graham Neubig. 2023. An in-depth look at gemini's language abilities. *arXiv preprint arXiv:2312.11444*.

Md Jahangir Alam, Ismail Hossain, Sai Puppala, and Sajedul Talukder. 2024. Advancements in multimodal social media post summarization: Integrating gpt-4 for enhanced understanding. In *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 1934–1940. IEEE.

Evaggelos Apostolidis, Eleni Adamantidou, Alexandros I. Metsai, Vasileios Mezaris, and Ioannis Patras. 2021. Video summarization using deep neural networks: A survey. *arXiv preprint arXiv:2101.06072*.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Harrison Blake. 2025. Multilingual capabilities of gpt-4 and llama.

Jessica Y Bo, Sophia Wan, and Ashton Anderson. 2025. To rely or not to rely? evaluating interventions for appropriate reliance on large language models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–23.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, and 1 others. 2005. The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. 2013. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2634–2641.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DR JIMBO A FUENTES. *" Influence of Price and Product Quality on Dining Preferences and On and Off Campus Food Choices of 4th-Year Marketing Students at a Private University.* Ph.D. thesis, Xavier University.

Gayatree Ganu, Noemie Elhadad, and Amélie Marian. 2009. Beyond the stars: Improving rating predictions using review text content. In *WebDB*, volume 9, pages 1–6.

Spandana Gella, Mike Lewis, and Marcus Rohrbach. 2018. A dataset for telling the stories of social media videos. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 968–974.

Toan Nguyen Hai, Ha Nguyen Viet, Truong Quan Xuan, and Duc Do Minh. 2025. A vietnamese dataset for text segmentation and multiple choices reading comprehension. *arXiv preprint arXiv:2506.15978*.

Gabriel Huang, Bo Pang, Zhenhai Zhu, Clara Rivera, and Radu Soricut. 2020. Multimodal pretraining for dense video captioning. *arXiv preprint arXiv:2011.11760*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

Ngoc C Lê, Nguyen The Lam, Son Hong Nguyen, and Duc Thanh Nguyen. 2020. On vietnamese sentiment analysis: a transfer learning method. In *2020 RIVF international conference on computing and communication technologies (RIVF)*, pages 1–5. IEEE.

The Anh Le and Hai Son Le. 2025. Latvis: Large-scale task-specific language model for low-resource vietnamese multi-document summarization. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(5):1–19.

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

Haixu Liu, Wenning Wang, Haoxiang Zheng, Penghao Jiang, Qirui Wang, Ruiqing Yan, and Qiuzhuang Sun. 2025. Multi-modal video feature extraction for popularity prediction. *Preprint*, arXiv:2501.01422.

Hui Liu and Xiaojun Wan. 2023. Models see hallucinations: Evaluating the factuality in video captioning. *arXiv preprint arXiv:2303.02961*.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52.

Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. 2022. Learning audio-video modalities from image captions. In *European Conference on Computer Vision*, pages 407–426. Springer.

Bang Nguyen, Van-Ho Nguyen, and Thanh Ho. 2021. Sentiment analysis of customer feedbacks in online food ordering services. *Business Systems Research: International journal of the Society for Advancing Innovation and Research in Economy*, 12(2):46–59.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. Phobert: Pre-trained language models for vietnamese. *arXiv preprint arXiv:2003.00744*.

Quoc-Nam Nguyen, Thang Chau Phan, Duc-Vu Nguyen, and Kiet Van Nguyen. 2023. Visobert: A pre-trained language model for vietnamese social media text processing. *arXiv preprint arXiv:2310.11166*.

Truc Mai-Thanh Nguyen, Dat Minh Nguyen, Son T Luu, and Kiet Van Nguyen. 2025. Vimrhp: A vietnamese benchmark dataset for multimodal review helpfulness prediction via human-ai collaborative annotation. In *International Conference on Applications of Natural Language to Information Systems*, pages 291–305. Springer.

Mayu Otani, Yale Song, Yang Wang, and 1 others. 2022. Video summarization overview. *Foundations and Trends® in Computer Graphics and Vision*, 13(4):284–335.

Minsu Park, Mor Naaman, and Jonah Berger. 2016. A data-driven study of view duration on youtube. In *Proceedings of the international AAAI conference on web and social media*, volume 10, pages 651–654.

Jielin Qiu, Jiacheng Zhu, William Han, Aditesh Kumar, Karthik Mittal, Claire Jin, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Ding Zhao, and 1 others. 2024. Mmsum: A dataset for multimodal summarization and thumbnail generation of videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21909–21921.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36.

Listia Rini, Joachim Jietse Schouteten, Ilona Faber, Michael Bom Frøst, Federico JA Perez-Cueto, and Hans De Steur. 2024. Social media and food consumer behavior: A systematic review. *Trends in Food Science & Technology*, 143:104290.

Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. 2014. Coherent multi-sentence video description with variable level of detail. In *German conference on pattern recognition*, pages 184–195. Springer.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Chi Tran and Huong Le Thanh. 2024. Lavy: Vietnamese multimodal large language model. *arXiv preprint arXiv:2404.07922*.

Phi Hung Truong and Anh Dao Kim. 2023. The influence of tiktok on young generation in vietnam. In *European Conference on Social Media*.

Caroline Violot, Tuğrulcan Elmas, Igor Bilogrevic, and Mathias Humbert. 2024. Shorts vs. regular videos on youtube: A comparative analysis of user engagement and content creation trends. In *Proceedings of the 16th ACM Web Science Conference*, pages 213–223.

Georg Von Krogh, Sebastian Spaeth, and Karim R Lakhani. 2003. Community, joining, and specialization in open source software innovation: a case study. *Research policy*, 32(7):1217–1241.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Christian Willmes, Daniel Kürner, and Georg Bareth. 2014. Building research data management infrastructure using open source software. *Transactions in GIS*, 18(4):496–509.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Jianhao Yan, Pingchuan Yan, Yulong Chen, Judy Li, Xianchao Zhu, and Yue Zhang. 2024. Gpt-4 vs. human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels. *arXiv preprint arXiv:2407.03658*.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, and 1 others. 2025. Qwen2. 5-1m technical report. *arXiv preprint arXiv:2501.15383*.

Mingyue Yuan, Jieshan Chen, Zhenchang Xing, Gelareh Mohammadi, and Aaron Quigley. 2025. A case study of scalable content annotation using multi-llm consensus and human review. *arXiv preprint arXiv:2503.17620*.

Savvas Zannettou, Olivia Nemes-Nemeth, Oshrat Ayalon, Angelica Goetzen, Krishna P Gummadi, Elissa M Redmiles, and Franziska Roesner. 2024. Analyzing user engagement with tiktok's short format video recommendations using data donations. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Shengyu Zhang, Ziqi Tan, Jin Yu, Zhou Zhao, Kun Kuang, Jie Liu, Jingren Zhou, Hongxia Yang, and Fei Wu. 2020. Poet: Product-oriented video captioner for e-commerce. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1292–1301.

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

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197*.

Luowei Zhou, Chenliang Xu, and Jason Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Appendix

A.1 Human Verification

Because the process uses LLMs, quality control is very important. In this workflow, humans with undergraduate-level education check after each step with LLM support for factual accuracy and contextual consistency, ensuring high-quality Human–LLM collaboration. First, the humans fix spelling mistakes and add missing details to the transcript that was first checked by GPT-4. Then, after Gemini produces a short description, the humans check again for spelling and meaning. Finally, the humans confirm the translation, which was also checked by GPT-4, to make sure the dataset is correct. Each checked sample costs $0.20.

Humans participate in the process to review and edit outputs generated by LLM models. They check for spelling errors, mistakes from transcripts, misinterpretation of context by the models, and any content produced by AI that is inappropriate or inconsistent.

A.2 Prompt Templates and Model Configurations

This subsection provides the full prompt templates and model configurations applied in each stage of the dataset construction pipeline. These prompts define the exact instructions given to the LLMs, including content extraction, description generation, translation, and bilingual verification. All configurations and constraints are preserved to ensure reproducibility and maintain consistency across the entire process.

- **Script Refinement:** LLM-generated scripts (from two model: GPT and Gemini) are cleaned by human annotators—removing off-topic content, fixing errors, and adding missing details from video frames or audio.
- **Description Creation:** Summaries must retain dish, price, and location details; exclude unrelated personal opinions.
- **Translation:** Vietnamese descriptions are translated into English by LLMs, with human checks to ensure tone, meaning, and cultural appropriateness.

Listing 1: Prompt templates and configurations for each stage.

```
Stage 1: GPT  Content Extraction
--------------------------------
Given the video script content, extract and fill
     in the following structured fields using
    only the information explicitly present in
    the script. Do not infer or add any
    information beyond what is given. Correct
    any spelling errors if necessary.

- Person:

- Location (Where):

- Address:

- Cooking Method:

- Price:

- Review Sentiment:

- Review Elements (e.g., quality, taste, service
    ):

* Rules:

- Use only information available in the script.

- Correct spelling errors when needed.

- Do not add or infer any additional content.

- All output must be written in Vietnamese.

Stage 2: Gemini  Description Generation
---------------------------------------
Given:
(1) The food review content;
(2) Supplementary viewer information, which may
    be edited or expanded if incomplete or
    inaccurate.

Task: Merge these two inputs into a single,
    concise description for a food review video.
     The description must remain faithful to the
     provided content, without adding unrelated
    information, and should retain all key
    details.

Additional requirement:
The final output must be written entirely in
    Vietnamese.

Stage 3: VinAI  Translation (vi - en)
-------------------------------------
Model: vinai/vinai-translate-vi2en-v2
Source language: vi_VN
Target language: en_XX
Beam size: 5
Max length: 1024

Stage 4: GPT  Bilingual Verification
------------------------------------
Act as a professional bilingual translation
    reviewer.
Compare the Vietnamese source and English
    translation.
If accurate, return as-is; if not, return
    corrected English version.
Rules: No explanation, no commentary, no extra
    text.
```

The exact prompt templates and model configurations used at each stage of the pipeline are shown in **Listing 1**. The pseudo-code format keeps the original wording, constraints, and parameter settings. This makes it possible to reproduce the process exactly as designed.

A.3 Example Data Samples

Sample records from the dataset are shown in **Figure 13**. They include the YouTube video URL, the original Vietnamese transcript, the human-polished Vietnamese description, and the verified English translation.

The first column holds the original YouTube video URLs. Next, the second column captures the transcripts automatically extracted from these videos. These transcripts are then refined into concise Vietnamese descriptions found in the third column, which are carefully polished by human editors. Finally, the fourth column features English summaries that have gone through thorough verification combining AI assistance and human review. Altogether, these steps show the journey from raw video content to a well-crafted bilingual dataset.

In addition to the main columns, the dataset also includes two others: one containing the refined transcript based on original script, and another holding rough English translations of the Vietnamese descriptions. Full details on all columns are provided in the linked in our GitHub repository to help researchers understand and use the dataset effectively.

A.4 Base models version

In our experiments, we employ four open-source LLMs. The selected models include:

- **Gemma 7B** (https://huggingface.co/google/gemma-7b)
- **Qwen2.5-7B-Instruct-1M** (https://huggingface.co/Qwen/Qwen2.5-7B-Instruct-1M)
- **DeepSeek-R1-Distill-Qwen-14B** (https://huggingface.co/avoroshilov/DeepSeek-R1-Distill-Qwen-14B-GPTQ_4bit-128g)

ID Video	Original Transcript	VI Description	EN Description
FOySxDUh4yg	tiếp tục cái series quán cơm bình dân mà giá không hề bình dân ở trên phố rồi Hôm nay tôi giới thiệu anh em đến một cái quán cơm khá là nổi tiếng xung quanh đây đó chính là cái quán cơm tên là Vinh thu	Chàng trai reviewer đã ghé thăm quán cơm Vinh Thu trên phố, một địa chỉ khá nổi tiếng trong khu vực. Anh đã gọi một số món đại diện như thịt kho, chả lá lốt, chả dươi, canh khoai và đậu tẩm hành.	The reviewer visited Vinh Thu rice restaurant on the street, a fairly famous place in the area. He ordered some representative dishes such as braised pork, grilled betel leaf pork roll, fried fish cake, potato soup, and fried tofu with scallions. However, he was quite disappointed with some dishes: the grilled betel leaf pork roll had leaves that did not fully wrap the meat, while the fried tofu with scallions lacked the scallion flavor.
kpZrL-rEzRQ	làm cái clip để nhắc nhở cho anh em ở Hà Nội ngoài bún xôi miến phở ra vẫn còn một cái món nó ngon điên cả lên đó chính là cái món trứng chén nướng cái món trứng chén nướng này ở Hà Nội không phải là dễ bắt gặp đâu Hoặc là nó có nhiều một khu nào đấy Tôi không biết tôi cứ giới thiệu cho anh em một cái hàng ở trên phố để tiện đi chơi về ghé ngang ăn chính ra cái món này nó sẽ rất rất là bình thường nếu mà không có cái sốt me đi kèm này một cái chén trứng với cả ba bốn quả trứng chim cút cùng với cả hành khô với cả ruốc nó sẽ rất là ngấy Nếu không có một tí chua chua của sốt me	Chàng trai reviewer muốn nhắc nhở mọi người về món trứng chén nướng ngon "điên cả lên" tại Hà Nội, và giới thiệu một quán ở 32 Hội Vũ, Hoàn Kiếm, Hà Nội. Anh chàng cho rằng món trứng chén nướng sẽ rất bình thường nếu không có sốt me đi kèm. Một chén trứng gồm ba bốn quả trứng chim cút cùng với hành khô và ruốc sẽ rất ngấy nếu thiếu vị chua của sốt me. Ngoài ra, anh cũng thử món trông như mực chiên nhưng hóa ra là bánh gà. Tất cả các món ở đây đều có thể ăn hoặc chấm cùng với sốt me.	The reviewer wants to remind everyone about the incredibly delicious grilled quail eggs in a cup in Hanoi, and recommends a spot at 32 Hoi Vu Street, Hoan Kiem District, Hanoi. He believes that grilled quail eggs would be quite ordinary without the accompanying tamarind sauce. A cup containing three to four quail eggs, along with fried shallots and shredded pork, would be too rich without the sour note of the tamarind sauce. Additionally, he tried a dish that looked like fried squid but turned out to be chicken fritters. All the dishes here can be eaten with or dipped in tamarind sauce.

Figure 11: Example dataset entries

- **Phi-4** (https://huggingface.co/microsoft/phi-4)

A.5 Example Base model Error - VietNamese and English

This subsection presents a representative example of an error generated by the baseline model when processing Vietnamese text. The example is selected to illustrate common challenges the model faces, such as handling tonal marks, word segmentation, and semantic ambiguity in Vietnamese language processing. By analyzing this case, we can better understand the limitations of the baseline model and identify potential directions for improvement. The detailed input, model output, and expected output are provided to highlight the nature of the error and its impact on the overall performance.

Model	Original Transcript	VI Description	EN Description
Gemma	chú này bán hủ tiếu là hổi lúc mình chưa đẻ rồi nữa mà cái tô hủ tiếu bờ viên ở đây ngon thiệt sự luôn á chú là người hoa mà đỗ ăn người Hoa là số vách rồi đó bán là có cái xe trước cửa nhà vậy nè vô ăn là bước vô nhà luôn vậy đó thấy được toàn cảnh ở bên trong luôn ở đây thì chú bán mỗi hủ tiếu thôi sẽ có là khô hoặc nước thì một tô là 35.000đ Á trong đó sẽ có ba viên bò viên nha bò viên thấy cũng ngon hủ tiếu thấy cũng ngon mà cái phần nước lèo thì chú làm thấy cũng ngon nữa Đúng là kinh nghiệm lâu năm có khác thiệt thì trong này có ba cái bèn à Ba cặp mà vô đây cái là hết chỗ rồi đó Nhưng mà vậy đó mà người ta ăn cũng lai rai quá trời quá đất khẻ thì gọi con to kho với tô nước cái phần nước lèo chú nêm nếm đậm đà thơm ngon dễ sợ bò viên thì chú nói không có bỏ vô cái nồi nước lèo luôn mà khi nào khách ăn thì chú mới bỏ vô điểm cộng đ là tương đen với sa tế đây cũng rất là ngon sền sệt cay cay cái này cũng không được gọi là cái quán nữa chỉ có cái xe Thôi chú làm theo kiểu người hoa hồi xưa á Ổ đây nằm tuốc trong hẻm Á địa chỉ là 22/53 Hà Tôn Quyền không Gình Gình đi ăn trễ xíu là hết trơn Ráng chịu đó viên bò viên thì cũng rất là bự và to ha ai mà muốn gọi thêm là là 5000 một viên rồi á tranh thủ ghé Đi ngon á nha.	chú này bán hủ tiếu là hổi lúc mình chưa đẻ rồi nữa	- The broth is very tasty. - The beef ball is also very delicious. - I really recommend this place. - The broth is made with 3 beef balls. - When you eat here, you will definitely be full. - The address is 22/53 Hà Tôn Quyền, Gình G
DeepSeek		Hủ tiếu là món ăn yêu thích của nhiều người, khi đến quán này, bạn sẽ được thưởng thức hủ tiếu bò viên với nước lèo đậm đà, thơm ngon. Giá cả hợp lý, chỉ 35.000 đồng cho một tô, trong đó có ba viên bò viên to và ngon. Quán nằm trong hẻm, có địa chỉ 22/53 Hà Tôn Quyền, bạn đến muộn một chút là hết sạch.	Hủ tiếu is a favorite dish for many people. At this eatery, you can enjoy a bowl of beef meatball rice noodles with a rich and fragrant broth. The dish comes with three large and tasty beef meatballs, all for a reasonable price of 35,000 VND per bowl. The restaurant is located in a small alley at 22/53 Hà Tôn Quyền Street. Be sure to arrive early, as it tends to sell out quickly.
Phi-4		Hủ tiếu của chú này có vị ngon và chất lượng tốt, đặc biệt là bò viên béo ngậy. Giá một tô hủ tiếu, bao gồm ba viên bò, chỉ 35.000 đồng, rất hợp lý với chất lượng thực phẩm. Mặc dù không phải là một quán ăn lớn, nhưng quán ăn di động này vẫn thu hút khách hàng với hương vị đặc trưng của ẩm thực Hoa và sự nồng nàn của nước lèo, tạo nên một trải nghiệm ẩm thực đáng nhớ	The noodle soop tastes great, with juicy beef meetbals. The price is resonable for the food qality.

Figure 12: Example dataset entries

Model	Original Transcript	VI Description	EN Description
Qwen	chú này bán hủ tiếu là hồi lúc mình chưa đẻ rồi nữa mà cái tô hủ tiếu bờ viên ở đây ngon thiệt sự luôn á chú là người hoa mà đồ ăn người Hoa là số vách rồi đó bán là có cái xe trước cửa nhà vậy nè vô ăn là bước vô nhà luôn vậy đó thấy được toàn cảnh ở bên trong luôn ở đây thì chú bán mỗi hủ tiếu thôi sẽ có là khô hoặc nước thì một tô là 35.000đ Á trong đó sẽ có ba viên bò viên nha bò viên thấy cũng ngon hủ tiếu thấy cũng ngon mà cái phần nước lèo thì chú làm thấy cũng ngon nữa Đúng là kinh nghiệm lâu năm có khác thiệt thì trong này có ba cái bèn à Ba cặp mà vô đây cái là hết chỗ rồi đó Nhưng mà vậy đó mà người ta ăn cũng lai rai quá trời quá đất khẻ thì gọi con to kho với tô nước cái phần nước lèo chú nêm nếm đậm đà thơm ngon dễ sợ bò viên thì chú nói không có bỏ vô cái nồi nước lèo luôn mà khi nào khách ăn thì chú mới bỏ vô điểm cộng đ là tương đen với sa tế đây cũng rất là ngon sền sệt cay cay cái này cũng không được gọi là cái quán nữa chỉ có cái xe Thôi chú làm theo kiểu người hoa hồi xưa á Ở đây nằm tuốc trong hẻm Á địa chỉ là 22/53 Hà Tôn Quyền không Gình Gình đi ăn trễ xíu là hết trơn Ráng chịu đó viên bò viên thì cũng rất là bự và to ha ai mà muốn gọi thêm là là 5000 một viên rồi á tranh thủ ghé Đi ngon á nha	chú này bán hủ tiếu là hồi lúc mình chưa đẻ rồi nữa	The noodle soup here is truly delicious, with flavorful beef dumplings and rich broth, but it's a bit pricey for the quality. The owner has been selling for a long time and has plenty of experience, but the location is a bit hard to find. A bowl of hủ tiếu here costs 35,000越南 段越南文翻成中文: Noodle soup ở đây ngon thật sự, phần bò viên và nước dùng đều rất đậm đà, nhưng giá hơi cao so với chất lượng. Chủ quán bán từ lâu nên có kinh nghiệm chế biến, nhưng vị trí hơi khó tìm. Một tô hủ tiếu ở đây có giá 35.000 đồng, gồm bò viên và nước dùng ngon miệng, tuy nhiên vị trí quán khó tìm và giá cả hơi cao so với chất lượng. Chủ quán có nhiều năm kinh nghiệm, nhưng vì chỉ bán từ xe đẩy nên không gian hơi chật. 中文：这里的面条汤确实很好吃，牛肉丸和汤头都很浓郁，但价格稍微贵了一些。店主经营了很久，经验丰富，但是店的位置不太好找。一碗河粉的价格是35

Figure 13: Example dataset entries

Number of Videos by Duration

Price	Count
0-20	5
0-40	234
0-60	1694
0-120	1999
0-180	2020

Figure 14: Number of Videos By Durations

Number of Videos by View

View	Count
Under 10.000	853
[10.000; 100.000]	557
[100.000; 1.000.000]	482
Over 1.000.000	128

Figure 15: Number of Videos By Views

Number of Videos by Year

Year	Count
2020	2
2021	110
2022	234
2023	539
2024	496
2025	639

Figure 16: Number of Videos By Year

Number of Videos by Food

Food_name	Count
Phở bò	140
Cá viên chiên	112
Mực chiên	38
Bún bò	32
Bánh trứng muối	32

Figure 17: Number of Videos By Food

Number of Videos by Location

Location	Count
Thành phố Hồ Chí Minh	1723
Thủ đô Hà Nội	286
Tỉnh Tuyên Quang	101
Tỉnh Ninh Bình	91
Thành phố Đà Lạt	37
Tỉnh Thừa Thiên - Huế	31
Tỉnh Hải Dương	13
Tỉnh Bà Rịa - Vũng Tàu	3
Tỉnh Kiên Giang	2
Tỉnh Biên Hòa	2
Tỉnh Bến Tre	2
Tỉnh Phú Yên	2
Tỉnh Quảng Ninh	2
Tỉnh Tiền Giang	1
Thành phố Đà Nẵng	1

Figure 18: Number of Videos By Locations

Number of Videos by Price

Price	Count
Under 100.000 VND	3540
[100.000VND; 300.000VND]	364
Over 300.000	62

Figure 19: Number of Videos By Price

A.6 Quantities of Diversity Analysis

This section presents a series of tables that illustrate the diversity of the dataset. Each table provides a statistical overview of the number of videos according to different criteria, such as price ranges, categories, or other relevant attributes. By summarizing these key indicators, the tables help users quickly grasp patterns, trends, and the distribution of videos within the dataset.