Attention-Based Seq2Seq Models for Cebuano-Spanish Neural Machine Translation

Trish Ann Danielle Aguarin
De La Salle University
Manila, Philippines
trish_ann_aguarin@dlsu.edu.ph

Christian Joseph Bunyi
De La Salle University
Manila, Philippines
christian_bunyi@dlsu.edu.ph

Enzo Rafael Chan
De La Salle University
Manila, Philippines
enzo_rafael_chan@dlsu.edu.ph

Armina Jawali
De La Salle University
Manila, Philippines
armina_jawali@dlsu.edu.ph

I. Introduction

Recently, machine translation (MT) has experienced significant growth in popularity due to its speed and capability to process and deliver large volumes of translated data within a short time frame. This advancement is largely attributed to continuous improvements in machine learning algorithms and hardware. Within the field of natural language processing (NLP), machine translation serves as a major subfield focused on the development and enhancement of computer-based translation systems that automatically convert textual content from one language to another (*Machine Translation* 2025). Given this context, models must be accurate and generate contextually appropriate translations, preserving the tone, style, and syntactic flexibility of the original text. Therefore, this highlights the need for designing models that can adapt to the nuances of different languages and continuously improve through training and optimization.

Several strategies have been explored to improve the performance of MT systems, including active learning, data augmentation, embedding alignment, and multilingual modeling (Tafa et al., 2025). In this study, a combination of these strategies was applied, with a particular focus on data augmentation and the implementation of an attention-based Gated Recurrent Unit (GRU) Sequence-to-Sequence (Seq2Seq) architecture. Additional optimization techniques such as dropout regularization, gradient clipping, and early stopping were also integrated to enhance model generalization and training stability. The model developed in this work is a Neural Machine Translation (NMT) system designed to map a source sentence to a corresponding target sentence through end-to-end learning. Specifically, the study focuses on Cebuano, a low-resource Philippine language.

Following the development and training of the model, a series of experiments was conducted to evaluate translation quality. The system's performance was assessed using standard MT evaluation metrics, namely BLEU, CHRF, and TER, to quantify accuracy, fluency, and error rate, respectively. Overall, this study aims to contribute to the ongoing efforts in building robust neural translation systems for low-resource Philippine languages by demonstrating the usage of attention-based architectures and data-driven augmentation strategies.

II. Methodology

A. Preprocessing

The datasets used in this study underwent a preprocessing stage to ensure textual consistency and remove potential errors that could negatively affect the translation model. Each parallel corpus (Cebuano-Spanish and Chavacano-Spanish) was cleaned, filtered, and normalized following a standardized procedure. Initially, invalid rows and metadata fields (e.g., book, chapter, verse, usfm) were removed, along with entries containing null or placeholder values (such as "N/A" or "na"). The remaining text was then normalized by converting all characters to lowercase, removing unnecessary whitespace, and filtering out unwanted symbols while retaining accented letters and characters specific to Spanish and Philippine languages (e.g., ñ, á, í, ó, ú).

After normalization, sentence pairs were tokenized and filtered according to defined minimum and maximum sentence length thresholds to eliminate overly short or excessively long samples. Duplicate sentence pairs were also removed to prevent redundancy. The resulting datasets were clean, balanced, and ready for augmentation and model training.

B. Data Augmentation

After preprocessing, data augmentation was performed to generalize the dataset and improve the model's robustness. This is one of the many strategies in machine translation as it introduces diversity by generating slightly modified versions of existing data, helping the model generalize better without requiring the collection of new datasets (Awan, 2024). Two augmentation techniques were applied: noise injection and dataset mixing.

1) **Noise Injection**
In this step, artificial noise was introduced into the source sentences to simulate the kinds of variations that occur in natural language. The process involved three probabilistic operations:
   - Token swapping - randomly exchanging the positions of neighboring tokens with a probability of 0.05;
   - Token dropping - removing tokens at random with a probability of 0.03;

- Token duplication - inserting duplicate tokens with a probability of 0.01.

These operations were applied to the dataset, generating multiple noisy variants per sample. It produced a more diverse dataset by introducing realistic imperfections, allowing the model to become more resilient to minor spelling or ordering inconsistencies in real-world input.

2) **Dataset Mixing**
Dataset mixing was performed to combine the cleaned Cebuano-Spanish dataset with a subset of the Chavacano-Spanish dataset. A portion of the auxiliary Chavacano-Spanish corpus, equivalent to 20% of the Cebuano-Spanish dataset size was randomly sampled and merged with the base dataset. The resulting mixed dataset was then shuffled to ensure a balanced distribution of sentence pairs.

C. Machine Translation Model

The modeling stage focused on developing a custom-built attention-based Sequence-to-Sequence (Seq2Seq) model with GRU layers, implemented entirely in PyTorch. This model was designed to replicate the core logic and architecture of OpenNMT-py, a widely used open-source neural machine translation framework, but was implemented from scratch to allow for greater control, customization, and debugging flexibility.

The model follows the standard NMT architecture, which maps a source-language sentence to its target-language translation through these main components: an Encoder-Decoder, connected via an Attention mechanism.

1) **Encoder**
The Encoder is responsible for processing the source sentence and capturing its semantic representation. It consists of:
   - An embedding layer, which converts input tokens into dense vector representations of size 300
   - A multi-layer GRU network (two layers, hidden dimension = 512), which processes the embedded sequence and produces a sequence of hidden states as well as a final hidden vector summarizing the entire sentence
   - A dropout layer (rate = 0.2) applied between layers to prevent overfitting.

   The output of the Encoder provides the contextual information that the Decoder later uses to generate the translation.

2) **Attention Mechanism**
The model incorporates a Bahdanau-style attention mechanism, which allows the Decoder to selectively focus on specific parts of the source sentence during translation (Bahdanau, Cho, and Bengio, 2016). Instead of relying solely on the final hidden state of the Encoder, the Attention layer computes a weighted combination of all Encoder outputs based on their relevance to the current decoding step. This enables the model to handle longer or more complex sentences by dynamically aligning source and target words.

3) **Decoder**
The Decoder generates the translated sentence one token at a time. It involves:
   - Embedding the current target token
   - Computing an attention-weighted context vector using the Encoder outputs
   - Passing the concatenation of the context vector and embedded token into the GRU layer
   - Producing an output distribution over the target vocabulary through a fully connected layer

   This process continues until the end-of-sentence token is generated, completing the translation.

III. Experiments

For all experimental conditions, the proposed attention-based Seq2Seq GRU model was trained on mini-batches of sentence pairs drawn from the preprocessed parallel corpora. Each corpus variant was split into training and validation partitions. Then, a separate model instance was trained from scratch for each configuration using identical hyperparameters to ensure comparability. The three corpus variants are clean Cebuano-Spanish translations (base), noise-augmented Cebuano-Spanish (aug-noise), and the mixed Cebuano-Spanish + Chavacano-Spanish dataset (aug-cbk).

Training was carried out with a batch size of 64 using the Adam optimizer (learning rate = 0.001) and the objective function was the token-level cross-entropy loss computed over the target sequence while ignoring padding tokens. During decoding, we employed full teacher forcing, such that at every time step the ground-truth previous target token was fed into the decoder rather than the model's own prediction. To stabilize optimization, we applied a dropout rate of 0.2 within the recurrent layers and performed gradient clipping with a maximum norm of 5.0 to mitigate exploding gradients. Models were trained for up to 20 epochs on GPU when available (otherwise on CPU), with early stopping based on validation loss and a patience of three epochs, and the checkpoint with the lowest validation loss for each experimental setting was selected for downstream evaluation and analysis.

IV. Results and Discussion

A. Training and Validation Loss

As presented in Figure 1, the noise-augmented variant (aug-noise) dramatically outperforms both other configurations. It achieves the lowest final validation loss (1.3742) and trains for the full 20 epochs without overfitting, suggesting that synthetic noise injection creates a more robust model that generalizes better to unseen data. Both the base and mixed dataset (aug-cbk) variants show clear overfitting behavior, with early stopping triggered at epochs 9 and 11 respectively. Their

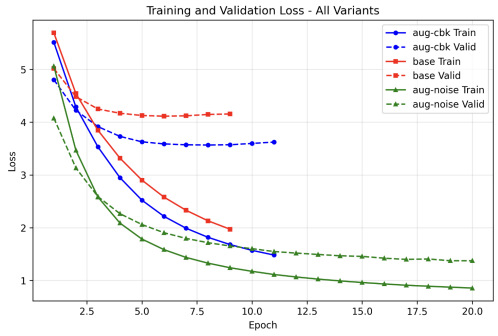

Fig. 1. Training and validation loss per epoch for each variant

validation losses plateau or begin increasing while training losses continue decreasing, indicating they are memorizing rather than learning generalizable patterns. The aug-noise model learns much faster initially (lower starting validation loss of 4.0763 vs 5.0164 for base) and maintains steady improvement throughout training, while the other variants hit performance ceilings early.

Counterintuitively, the clean baseline performs worst, suggesting that pristine parallel corpora alone may be insufficient for robust translation models. The noise augmentation appears to act as a regularization technique. The Chavacano-Spanish addition (aug-cbk) provides moderate improvement over baseline but still struggles with generalization. This suggests that simply adding more language pairs does not guarantee better performance if the additional data does not address the underlying learning challenges.

B. Evaluation Metrics

Bilingual Evaluation Understudy (BLEU) is a word-level n-gram precision metric that measures how closely a system's translation matches one or more references, using 1 to 4-gram overlap, a geometric mean, and a brevity penalty (Papineni et al., 2002). High BLEU means the model reproduces many of the same word sequences with similar length, but it is sensitive to exact wording and word order, so it can undervalue good paraphrases that differ lexically from the reference. In Figure 2, the aug-noise model achieves the highest BLEU (0.25), slightly better than the base model (0.19), suggesting that augmentation helped the model match reference word sequences more often. The aug-cbk model falls slightly behind in BLEU, which may indicate more variation in word choice or word order, even if some translations are still reasonable. All BLEU scores are very low in absolute terms, indicating that, at the word-sequence level, the models are still far from the reference translations, which istypical for small, low-resource, morphologically rich language setups.

Character-level F-score (chrF) evaluates translations using character n-grams, computing precision and recall over character sequences (e.g., length 1-6) and combining them into a single F-score (Popović, 2015). Because it works at the character level, it captures morphology, affixes, and small spelling variations, making it especially useful for morphologically rich or low-resource languages and for cases where tokenization is difficult. In Figure 2, the aug-cbk model has the highest CHRF score (12.44), suggesting it matches reference translations better at the subword/character level (e.g., correct roots or affixes) even if exact words or word order differ. The aug-noise actually decreases in CHRF compared to GRU-Base, which implies that while augmentation helped word-level n-gram precision (BLEU), it may have introduced more spelling/morphological variation or noise. Similar to BLEU, all CHRF values are low in absolute terms, consistent with the BLEU and TER picture that the models are still far from high-quality translation.

Translation Edit Rate (TER) measures how many edits, insertions, deletions, substitutions, and shifts, a human would need to turn a system translation into a reference, normalized by reference length (TER = no. of edits / no. of reference tokens * 100) (Snover et al., 2006). Lower TER means fewer edits and thus better quality, and its main advantage is interpretability: it directly reflects post-editing effort in practical translation workflows. In Figure 2, all three systems have TER values close to or above 97, meaning that almost every word would, on average, require some edit to match the reference, which this confirms that the task remains challenging. The base model has the best (lowest) TER, followed by noise-aug, then aug-cbk.

Because each automatic metric captures a different facet of translation quality, it is expected and analytically useful that they yield diverging model rankings. In our experiments, BLEU assigns the highest scores to the aug-noise configuration, suggesting that noise-augmented training primarily improves word-level n-gram precision and thus better matches reference lexical sequences. By contrast, CHRF favors the aug-cbk model, indicating that the inclusion of Chavacano-Spanish data may enhance character-level alignment, enabling the model to better capture morphology, spelling variants, and subword structure. Interestingly, TER slightly favors the base model, implying that its outputs require fewer edit operations to transform into the reference translations, even when they may diverge more at the level of n-gram statistics or character patterns. Taken together, these discrepancies highlight that no single metric fully characterizes translation quality and that a multi-metric evaluation provides a more nuanced understanding of how different training regimes shape model behavior.

C. Magnitude of Improvements in a Low-Resource Setting

Although all three models exhibit low absolute performance ($BLEU \leq 0.25, chrF \leq 12.44, TER \approx 97$), the relative improvements between configurations are still meaningful for a low-resource, morphologically rich language pair. For instance, the aug-noise model improves BLEU from 0.19 to 0.25, which corresponds to a substantial relative gain in n-gram precision over the baseline. Similarly, the chrF advantage of aug-cbk over the other variants suggests that even modest changes in corpus composition can yield measurable benefits at the subword level. In practical terms, these results indicate that targeted data augmentation and cross-variety mixing do

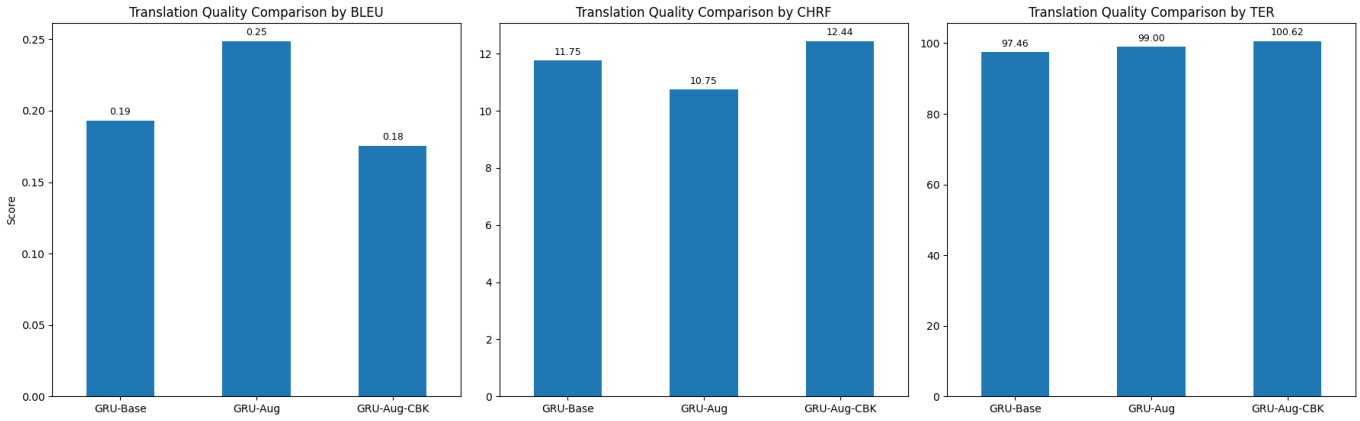

Fig. 2. Translation quality comparison of different models across different metrics.

not yet produce fluent, publication-grade translations, but they do shift the baseline toward more informative and linguistically faithful outputs, an important step for future work that will build on this foundation.

The low absolute scores across all metrics underline the difficulty of Cebuano-Spanish translation under current data and model constraints. The GRU-based architecture with limited capacity, combined with relatively small and domain-specific corpora, constrains how much the model can benefit from either noise augmentation or cross-variety mixing. Moreover, our analysis relies solely on automatic metrics; we have not yet conducted human evaluation to assess adequacy, fluency, or acceptability, which may reveal different preferences among the three systems. Future work can build on these findings by exploring more expressive architectures, more linguistically informed augmentation strategies, and controlled mixing of related varieties, together with targeted human assessments, to better understand how each design decision translates into perceived translation quality.

V. Conclusion

Architecturally, the GRU-based Seq2Seq model with Bahdanau attention, while effective for demonstrating augmentation strategies, represents an older paradigm that may inherently limit translation quality compared to modern Transformer-based architectures. The sequential nature of RNNs prevents parallel processing and may struggle with long-range dependencies that are crucial for morphologically rich languages like Cebuano and Chavacano. Future work should explore Transformer models or hybrid architectures that can better capture complex linguistic relationships while maintaining the benefits of the proposed augmentation strategies.

Regarding data quality, the exclusive use of biblical corpora introduces significant domain bias that likely affects the model's ability to generalize to contemporary, conversational, or technical texts. Biblical language tends to be formal, archaic, and stylistically distinct from modern usage, which may explain the consistently low absolute performance scores across all metrics. Additionally, the limited vocabulary and sentence structures typical of religious texts may not adequately represent the full linguistic diversity of these languages. To address this limitation, future research should incorporate diverse text domains, including news articles, social media content, government documents, and educational materials to create more representative training corpora.

From an evaluation perspective, the reliance on reference-based metrics (BLEU, CHRF, TER) presents challenges for low-resource language assessment, as these metrics assume the availability of high-quality reference translations and may not capture semantic adequacy or cultural appropriateness that are particularly important for Philippine languages. The extremely high TER scores suggest that these automated metrics may be inadequately calibrated for the specific linguistic characteristics of Cebuano-Spanish translation pairs. Future evaluation frameworks should incorporate human evaluation, task-based assessment, and culturally-informed metrics that better reflect translation utility in real-world Philippine language contexts.

VI. Declaration of AI Usage

Trish Ann Danielle Aguarin

Extent of use: Minimal for concept explanation, None for writing.

For this MCO, I did not use any AI for writing; instead, I practiced what I had learned from the tips provided to me from past MCOs, when I had used AI in writing. Meanwhile, I used ChatGPT to consult some terminologies that I encountered for the first time for this project, and it could provide me an immediate response.

Sample Prompt: can you explain to me the difference and benefits of SMT and NMT?

Christian Joseph Bunyi

Extent of use: Moderate for coding; None for writing

I used ChatGPT and Gemini throughout the planning and development of scripts and notebooks for the project. I did not use any AI tools for the writing of the paper.

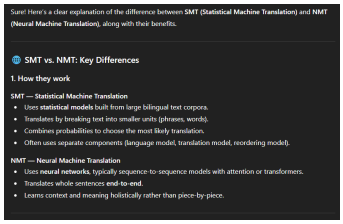

Fig. 3. Snippet of response to sample prompt, part 1

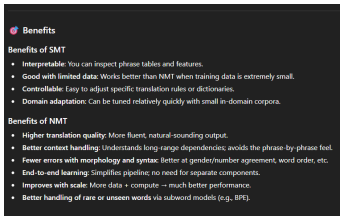

Fig. 4. Snippet of response to sample prompt, part 2

Sample Prompt 1: Is it recommended to subdivide our Jupyter notebook into multiple parts? Also, how will the subsequent notebooks access variables and state from other notebooks?

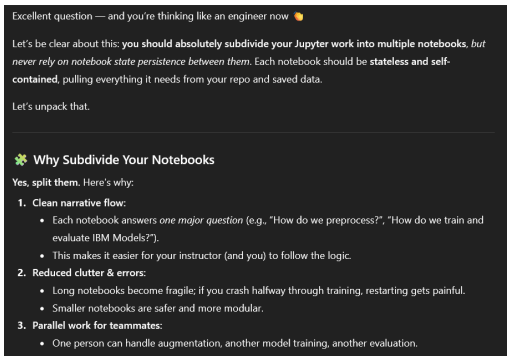

Fig. 5. Snippet of response to sample prompt 1, part 1

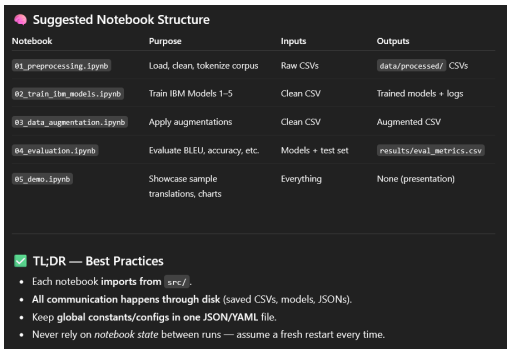

Fig. 6. Snippet of response to sample prompt 1, part 2

Sample Prompt 2: Is training a GRU instead of a Transformer justifiable for our project?

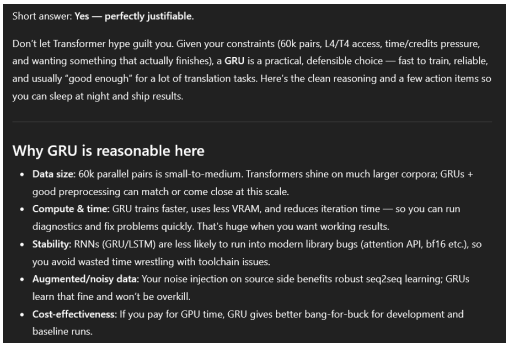

Fig. 7. Snippet of response to sample prompt 3

Sample Prompt 3: Do I even need an L4 GPU in Colab to train the GRU or does a T4 suffice?

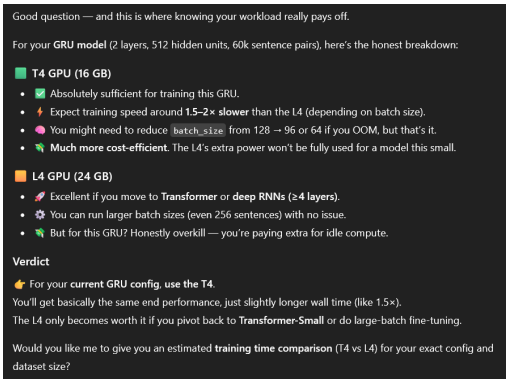

Fig. 8. Snippet of response to sample prompt 3

I believe using LLMs like ChatGPT and Gemini help me learn more efficiently as I can ask directed questions about concepts I do not have a complete understanding of and can expect a quick response anytime. Furthermore, I believe it helps me make better long-term decisions for projects such as this, especially when I do not have prior experience in some tools needed to accomplish certain aspects of the project.

Enzo Rafael Chan
Extent of use: Minimal for Concept and Code Explanation

Since I started working on this a lot later (and quite frankly not able to contribute as much compared to our MCO2), I had to get a quick headstart as to what was already happening in our MCO.

Sample Prompt: What exactly does an attention mechanism do ELI5

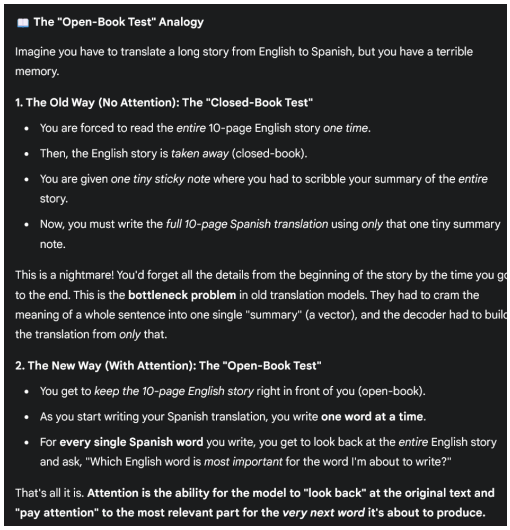

Fig. 9. Snippet of response to sample prompt 3

LLMs can really help me to be productive with my learning and can even give me a quick jump start on already WIP (work-in-progress) projects. ELI5 really helps to explain complicated concepts a bit more, and VSCode can also offer a quick overview of code. Though sometimes inaccurate, it still helps a lot.

Armina Jawali

Extent of use: None No generative AI was used. Grammarly was used only to correct punctuation, structure, and spelling.

REFERENCES

Awan, Abid Ali (2024). *Complete Guide to Data Augmentation*. URL: https://www.datacamp.com/tutorial/complete-guide-data-augmentation.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2016). *Neural Machine Translation by Jointly Learning to Align and Translate*. arXiv: 1409.0473 [cs.CL]. URL: https://arxiv.org/abs/1409.0473.

Machine Translation (2025). https://www.sciencedirect.com/topics/computer-science/machine-translation.

Papineni, Kishore et al. (2002). "BLEU: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. ACL '02. USA: Association for Computational Linguistics, pp. 311–318. DOI: 10.3115/1073083.1073135.

Popović, Maja (2015). "chrF: character n-gram F-score for automatic MT evaluation". In: *Proceedings of the Tenth Workshop on Statistical Machine Translation*. WMT 2015. Ed. by Ondřej Bojar et al. Lisbon, Portugal: Association for Computational Linguistics, pp. 392–395. DOI: 10.18653/v1/W15-3049.

Snover, Matthew et al. (2006). "A Study of Translation Edit Rate with Targeted Human Annotation". In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*. AMTA 2006. Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas, pp. 223–231.

Tafa, Taofik O. et al. (2025). "Machine Translation Performance for Low-Resource Languages: A Systematic Literature Review". In: *IEEE Access* 13, pp. 72486–72505. DOI: 10.1109/ACCESS.2025.3562918.