Multi-modal Neural Machine Translation for Low-Resource Classical Persian Poetry: A Culture-Aware Evaluation

Soheila Ansari, Mounir Boukadoum, Fatiha Sadat
Université du Québec à Montréal
{ansari.soheila, boukadoum.mounir, sadat.fatiha}@uqam.ca

Abstract

Persian poetry, particularly Rumi's *Masnavi-ye-Ma'navi*, is known for its complex form, mystical narrative style, rich cultural information, and linguistic nuances, and is considered a low-resource domain. Translating Persian poetry is a challenging task for neural machine translation (NMT) systems. To address this challenge, we present a novel multi-modal NMT system for Rumi's *Masnavi* in four stages. First, we built a new multi-modal parallel Persian-English corpus of 26,571 aligned verses from all six books of *Masnavi*, and all paired with aligned audio recitations. Second, a strong text-only baseline is developed by applying domain-adaptive fine-tuning to mBART-50, pre-trained on a large monolingual Persian poetry corpus, followed by training on the parallel *Masnavi* corpus (train set). Third, we extend this model to a multi-modal scenario by adding aligned audio representations using a cross-attention fusion mechanism. Fourth, we conduct a culture-aware evaluation. We propose a culture-specific item (CSI) evaluation approach by developing a CSI classification system and a Persian-English CSI dictionary alongside the standard MT metrics. Our findings demonstrate that integrating audio recitations increased the BLEU score from 9.85 to 17.95, and raised CSI-recall from 61.60% to 82.04%, suggesting greater consistency in producing culturally meaningful terms.

1 Introduction

Rumi's *Masnavi-ye-Ma'navi* has a central place in Persian cultural and literary heritage, representing a profound synthesis of theological and metaphorical discourse often referred to as the "Quran in Persian" (Sedaghat, 2020). This poetry is known for its mystical language, whose complexity often causes both human and machine translators to unintentionally impose ideological or structural biases that diminish the text's intentional indeterminacy (Sedaghat, 2020). Although neural machine translation (NMT) has made significant advancements, its application to classical poetry remains a demanding and largely unexplored area (Chakrabarty et al., 2021), (Ghassemiazghandi, 2023). Standard NMT models, typically trained on massive datasets like news and articles, frequently struggle with the low-resource nature of classical Persian and the dense figurative patterns found in the works of masters like Rumi (Ghassemiazghandi, 2023). As a result, they frequently fail to preserve the source text's distinctive "literary touch" and worldview (Ghassemiazghandi, 2023), (Chakrabarty et al., 2021). Research indicates that domain-adaptive fine-tuning on poetic corpora significantly outperform general-domain models, yet a critical gap remains: the lack of comprehensive, annotated datasets that capture the complex rhythmic and cultural rules of Persian language (Ghassemiazghandi, 2023), (Chakrabarty et al., 2021). A defining characteristic of Persian poetry is that it is intended to be heard rather than read silently. Recitation, through its unique rhythm, tone, and pauses, helps to resolve and clarify the ambiguities of the written Persian and reveal the poet's intent. In this context, multi-modal approaches that integrate audio features with text offer a game-changing ability for translation models, providing access to the aforementioned details that written words alone cannot convey (Shahrestani and Haghir Chehreghani, 2025). Furthermore, traditional evaluation metrics like BLEU often fail to track whether culture-specific items (CSIs), such as Sufi concepts, religious figures, or Quranic phrases are accurately preserved or erased during the translation process.

To address these limitations, this paper introduces, to the best of our knowledge, the first multi-modal NMT (MMNMT) model specifically designed for the *Masnavi*, supported by a recitation-aligned multi-modal Persian–English corpus. This dataset is made up of 26,571 aligned verse pairs

Proceedings of the First Workshop on NLP and LLMs for the Iranian Language Family (SilkRoadNLP), pages 131–139
March 29, 2026 ©2026 Association for Computational Linguistics

across all six books of *Masnavi*, accompanied by corresponding Persian audio recitations. A robust Persian poetry fine-tuned text-only baseline (PPFT) is established, and afterwards extended to a multi-modal model. We aim to better preserve the depth and artistic intent of Rumi's poetry after translation. Therefore, we propose a CSI-aware evaluation framework using a customized CSI lexicon to diagnose and enhance the cultural authenticity in machine translation outputs. Through this work, we demonstrate that multi-modal integration not only clarifies interpretive ambiguity but also substantially improves the accuracy of culturally grounded lexical realizations.

This paper is structured as follows: we present a review of the related work to our topic in Section 2. Section 3 is dedicated to describing our produced multi-modal dataset. In Section 4, we introduce our CSI taxonomy, annotation process, and the construction of a Persian–English CSI lexicon. Our proposed methodology is elaborated in Section 5. Section 6 presents results and CSI-aware evaluation and analysis. Finally, Section 7 discusses limitations and outlines directions for future work followed by concluding remarks.

2 Related Work

Recent studies outside the realm of Indo-European languages investigated on how using literary and poetic texts can serve as an important role to preserve and revitalize endangered Indigenous languages. For instance, Cadotte et al. (2024) demonstrated that when traditional bilingual data is scarce, literature and poetry can help to develop machine translation (MT) models, specifically for languages like Innu-Aimun. This study supports the fact that employing culturally rich texts is a practical approach for training specialized systems where generic data is insufficient.

Chakrabarty et al. (2021) showed improvements in neural poetry translation through multilingual fine-tuning on poetic corpora, emphasizing the preservation of poetic style and meaning. However, they evaluated their work by using standard metrics like BLEU and COMET. Such metrics struggle to preserve cultural aspects, which is the challenge our work aims to address by introducing a CSI lexicon for the evaluation phase. This enables a fair evaluation that takes into account the preservation of cultural weight throughout the translation process.

The evaluation of cultural fidelity in MT has recently moved toward tracking CSIs. Yao et al. (2024) introduced the culturally-aware machine translation (CAMT) corpus and the CSI-Match metric (a metric that determines how closely a model's output matches reference translations for CSIs, using a fuzzy matching approach) to assess how models handle cultural entities in modern prose, such as Wikipedia articles. Comparably, Thakur (2025) proposed culturally-grounded chain-of-thought (CG-CoT) for interpretative analysis of Yoruba proverbs using retrieval-augmented generation. Although these works highlight the inadequacy of standard metrics like BLEU for cultural tasks, they focus on modern domains or reasoning sequences (Yao et al., 2024), (Thakur, 2025). In another study, Sadr et al. (2025) found that AI models often fail to understand Persian social rules like Taarof because they are biased toward Western-style directness. This is known as a "cross-cultural pragmatics problem" (Stadler, 2012), in which the literal words used in a conversation are different from what the person actually means. Because these rituals feature a form of "polite verbal wrestling" (Rafiee, 1991), text-only models that are fixated on literal meanings cannot capture the true cultural intent. This emphasizes the importance of specialized tools to evaluate how well the cultural weight of Persian language is preserved during translation (Sadr et al., 2025). Our work aims to move beyond generic models by introducing a specialized Persian poetry fine-tuned architecture (PPFT) that addresses the linguistic challenges of classical Persian verse. As it will be elaborated in Section 5.1, we construct an 18-category CSI lexicon specifically for the *Masnavi*, providing a targeted "Recall" metric that measures the preservation of classical cultural weight.

Most current machine translation research typically relies on either image-text multi-modal approaches or text-only methods. This means that the integration of information extracted from speech has received comparatively less attention. Current multi-modal machine translation research is heavily biased toward image-text combinations for short captions. Parida et al. (2024) and Rajpoot et al. (2024) utilized visual context from images to resolve polysemy in languages like Hindi and Bengali. As noted in the survey by Lupascu et al. (2025), text-image pairs represent 63% of the literature, while audio-text integration for low-resource languages remains significantly unexplored.

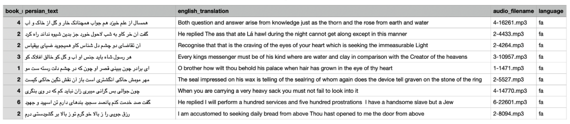
While Shahrestani and Haghir Chehreghani (2025) utilized deep learning for prosody recognition in Persian poetry, they mostly focused on classifying metrical patterns rather than translation. Xiang et al. (2025) empirically quantified the "modality gap" between speech and text inputs in large language models, identifying differences in semantic comprehension performance. Their research concentrates on conversational and synthetic speech in English. We aimed to addresses this gap by using self-supervised Persian speech encoders (Wav2Vec2) (Babu et al., 2021) and a cross-attention fusion mechanism (Vaswani et al., 2017) to capture essential audio features for classical poetry *Masnavi* that written text alone cannot provide.

Recent advancements in large language models (LLMs) have enabled new approaches to poetic interpretation. Gao et al. (2024) utilized ChatGPT to translate Chinese classical poetry, demonstrating that prompting models to interpret symbols or provide explanations first can enhance translation quality and readability. Although our work also employs LLMs (i.e., GPT-4) within a combined annotation framework, we distinguish our contribution by introducing a structured CSI-Recall metric as a diagnostic tool. This enables measurable evaluation of cultural transfer that goes beyond the subjective "Overall Impression" used in previous human evaluations (Wang et al., 2024).

Compared to the state of the art, our contribution is distinguished by two key aspects: the integration of recitation audio to clarify ambiguities, and the development of a CSI lexicon that enables evaluation of cultural fidelity, a dimension overlooked by conventional metrics such as BLEU and standard domain adaptation techniques.

3 Multi-modal Corpus of *Masnavi*

We construct and introduce a multi-modal Persian-English corpus derived from *Masnavi-ye-Ma'navi*. Its linguistic structure is characterized by "anomalous speech," where conventional grammatical patterns are intentionally altered to enhance aesthetic and emotional expression (Khanmohammadi et al., 2023). In addition, Persian poetry functions within a hybrid semiotic system that intertwines linguistic meaning with rhythmic, musical, and transcendental dimensions, further increasing abstraction and interpretive ambiguity (Sedaghat, 2020). These properties make *Masnavi* a challenging yet representative testbed for culturally grounded machine translation, particularly in low-resource settings. As discussed, this area remains underrepresented in existing MT benchmarks for Iranian languages.



book_	persian_text	english_translation	audio_filename	language
4	همسال از علم خیزد هم جواب همچنانک خار و گل از خاک و آب	Both question and answer arise from knowledge just as the thorn and the rose from earth and water	4-16261.mp3	fa
2	گفت آن خر کاو به شب لاحول خورد جز بدین شیوه نداند راه کرد	He replied The ass that ate Lá hawl during the night cannot get along except in this manner	2-4433.mp3	fa
2	آن تقاضای دو چشم دل شناس کاو همیجوید ضیای بیقیاس	Recognise that that is the craving of the eyes of your heart which is seeking the immeasurable Light	2-4264.mp3	fa
3	هر رسول شاه باید جنس او آب و گل کو خالق افلاک کو	Every kings messenger must be of his kind where are water and clay in comparison with the Creator of the heavens	3-10957.mp3	fa
1	ای برادر چون ببینی قصر او چون که در چشم دلت رسته ست مو	O brother how wilt thou behold his palace when hair has grown in the eye of thy heart	1-1471.mp3	fa
2	مهر مومش حاکی انگشتری است باز آن نقش نگین حاکی کیست	The seal impressed on his wax is telling of the sealring of whom again does the device tell graven on the stone of the ring	2-5527.mp3	fa
4	چون جوالی بس گرانی میبری زان نباید کم که در وی بنگری	When you are carrying a very heavy sack you must not fail to look into it	4-14770.mp3	fa
6	گفت صد خدمت کنم پانصد سجود بندهای دارم تن اسپید و جهود	He replied I will perform a hundred services and five hundred prostrations I have a handsome slave but a Jew	6-22601.mp3	fa
2	رزق جویی را ز بالا خو گرم تو ز بالا بر گشودستی درم	I am accustomed to seeking daily bread from above Thou hast opened to me the door from above	2-8094.mp3	fa

Figure 1: Partial representation of the produced multi-modal dataset.

3.1 Text Sources and Alignment

Persian text and English translations were collected from publicly available *Masnavi* repositories that provide full access to the Persian poems and their translations (Javadi et al., 2015), (Lewis, 2015), (Nicholson and Sufism.org, 2013). Different Persian literature repositories (e.g., Ganjoor [1]) are used for cross-checking verse boundaries and book structure. Each example is indexed and aligned at the verse unit. This indexing is also used to identify the corresponding audio file. The final prepared corpus consists of aligned Persian source verses, their English translations, and corresponding Persian audio recitations. A partial representation of the dataset structure is shown in Figure 1.

3.2 Audio Modality

Audio recitations were obtained from *Masnavi* audio repositories designed for verse-level listening (Javadi et al., 2015). We map each audio clip to its verse identifier and resample audio to a consistent sampling rate (16 kHz).

3.3 Corpus Statistics

As noted in Table 1, the corpus contains 26,571 aligned Persian-English verse pairs and 26,571 corresponding Persian audio files, utilizing all six books. We split the whole corpus into three sets of train (21,255), validation (2,658), and test (2,658). Each set includes: `book_id`, `persian_text`, `english_translation`, `audio_filename`, and `language`. Because literary translations may vary in literalness and poetic style, we treat the English side as a *reference translation* rather than a unique legitimate target and complement standard MT metrics with culturally targeted evaluation that is further discussed in Section 5 and Section 6.

[1] `https://ganjoor.net/moulavi/masnavi`

	Text	Audio
Size	26,571 lines/files	
Train	21,255 lines/files	
Validation	2,658 lines/files	
Test	2,658 lines/files	

Table 1: Statistics of multi-modal *Masnavi* corpus dataset and its division to train, validation, and test sets.

4 Cultural Aspect in Persian Poetry

The Persian language poses intrinsic challenges for NMT that extend well beyond lexical sparsity (Sedaghat, 2020). Such challenges would raise up in the translation of Persian classical poetry, particularly in low-resource settings. Prior work has noted the phenomenon of *second-degree deformation*, where translators (humans) or machines apply filters (e.g., secularization or Islamization biases) that reduce the poet's intent into fixed interpretations (Sedaghat, 2020). In other words, literal translation strategies are especially inadequate for this complex task. Moreover, generic NMT systems can generate false content or degrade metaphorical expressions upon processing figurative language (Chakrabarty et al., 2021).

Recent research suggests that addressing poetic translation requires architectural and training adaptations rather than mere scaling. Applying approaches augmented with domain-adaptive objectives can better respect poetic limitations than generic NMT systems (Khanmohammadi et al., 2023). However, we believe that tracking cultural-aware aspect of these translations by employing CSI-evaluation, can compensate for the inadequacy and deficiency of these methodologies.

To achieve this, we introduce the CSI lexicon for formalizing and evaluating cultural specificity in Persian–English poetic translation (detailed in Section 5).

5 Proposed Methodology

5.1 CSI Lexicon

To address the aforementioned limitations, this work models cultural fidelity in Persian–English poetry translation through CSIs. CSIs represent explicit mentions of culturally grounded entities and concepts, such as Sufi concepts, Quranic references, doctrine, and symbolic animals that hold meaning, semantic, and inter-textual weight beyond their literal expression. In contrast, standard MT evaluation metrics are poorly suited for this matter.

- Step 1: We developed a detailed label for the CSIs by integrating concepts from Persian literary studies and cultural linguistics. The taxonomy includes 18 culture-specific categories, listed as: person, place, Sufi concept, mystical phrase, supernatural being, number/color, medical/scientific concept, Quranic reference, religious concept, virtue, divine attribute, foreign Arabic/Turkic term, animal symbol, natural element, object symbol, sound/music, main doctrine of love, and other. Each label is accompanied by definition and some examples. These labels were manually refined and cleaned through iterative review by using cultural references from classical sources.

- Step 2: We annotated *Masnavi* verses with these CSI labels. In order to do that, we adopted a hybrid approach combining manual supervision as well as an LLM assistance (i.e, GPT-4). A unique prompting pipeline was used for the LLM to extend the annotations. Prompts provided the verse, the full taxonomy, and detailed instructions. The model predicted relevant CSI spans and assigned appropriate labels. All LLM-generated annotations were reviewed and corrected by the author to ensure high quality. The resulting dataset consists of 1,000 annotated verses that were randomly selected from the training set using a fixed seed for reproducibility.

- Step 3: The cleaned, filtered annotated file from the previous step is converted into token-level BIO tagging format (each token is marked as Beginning (B), Inside (I), or Outside (O) of a labeled span).

- Step 4: We trained a CSI tagger using our PPFT model and the BIO-tagged dataset.

- Step 5: We ran the trained tagger on the *Masnavi* corpus. Therefore, our whole corpus is tagged. As a result, we have fully tagged *Masnavi* corpus with CSI labels and spans.

The final phase is to build the Persian-English CSI lexicon, where each Persian span is mapped to its aligned English translation. For each pair, we collected all aligned English phrases across the corpus (train set) and aggregated them into a candidate set with noting the frequency and alignment confidence statistics. To improve precision and reduce noise, we performed several cleaning steps, such as canonicalization of English forms, frequency and rank filtering, and label consistency filtering.

After all these steps, we produced three lexicon versions denoted as strict, soft and broad. The first version, strict core lexicon, contains only high confidence spans that are verified by native human that serves as a gold-standard reference for evaluation. The second version, soft core lexicon, represents all tagged spans, that is a more relaxed version compared to the strict one, but still enforcing strong alignment confidence. The third version, broad lexicon, intends to maximize coverage. It includes all aligned CSI candidates above a minimal confidence threshold. It is noisier but still valuable for culture-aware evaluation.

We concluded that having three versions instead of a single lexicon was a better approach. A single version would risk being either too restrictive (i.e., missing valid translations) or non-restrictive by adding unwanted noise. This three-version approach enables us to demonstrate robustness across evaluation settings, and analyze the translation systems' response to progressively broader cultural criteria. It is worth mentioning that each lexicon is derived from the same alignment and tagging procedure, differing only in filtering thresholds over alignment confidence and frequency.

Finally, the lexicon offers substantial coverage of frequent and semantically significant CSIs. This structural approach allows the lexicon to serve as the following two objectives: either as a culture-aware evaluation framework, or as an analytical resource for examining cross-cultural transfer phenomena in automated translation systems. In this work, we employed the generated CSI lexicon to support the culture-aware evaluation for our machine translation models.

5.2 Multi-modal NMT system for Persian Poetry

While our work focuses on a neural framework, the choice between statistical machine translation (SMT) and NMT in low-resource settings continues to be a subject of discussion. Cadotte et al. (2024) found that for extremely limited datasets of fewer than 4,000 sentences, SMT outperformed NMT. However, because our *Masnavi* corpus consists of over 26,000 verse pairs, we can overcome such limitations and utilize a domain-adaptive neural baseline (denoted as PPFT) that is further enhanced by employing the audio modality.

The overview of our proposed multi-modal NMT framework is illustrated in Figure 2. After the preprocessing phase as discussed earlier, we implemented and evaluated a strong text-only unimodal baseline (denoted as the second phase in Figure 2). It is worth mentioning that almost all of general NMT systems are trained on the general domain of Persian language like news, and Wikipedia. Therefore, effective poetic translation requires changes in model architecture and training strategy. One way to have a more strong and meaningful translation is to have a strong baseline. However, culturally-based translation challenges, especially in low-resource settings, continue to persist even when using a stronger text-only baseline. In this regard, we pre-trained and fine-tuned the mBART50 (Tang et al., 2020) by using Persian poetry. We used a large monolingual Persian poetry corpus containing 1M lines, that is publicly available [2]. Then, we fine-tuned it on our parallel *Masnavi* Persian-English corpus (train set), denoted as Persian poetry fine-tuned text-only model (PPFT). We further used this model as the base for the rest of the multi-modal models.

As it is shown in Figure 2, the third phase regards developing the multi-modal NMT. This model is defined by the interaction between two high-dimensional latent spaces: the audio space (Wav2Vec 2.0) and the semantic poetry space (PPFT). In this system, two distinct encoders are employed. The audio encoder (Wav2Vec 2.0 XLS-R) (Babu et al., 2021) extracts phonetic representations, and the text encoder (PPFT), processes the Persian verses. We implemented a multi-modal fusion layer using a multi-head cross-attention mechanism (Vaswani et al., 2017). In result, the model can identify phonetic details in the Masnavi recitation to the related words or meaning in the verse. The resulting multi-model model is called MMNMT model. Finally, fused multi-modal representation is converted into a target English sequence by using the mBART50 decoder (Tang et al., 2020).

6 Evaluation Framework

Tables 2 and 3 respectively report the results of the NMT/MMNMT models and the CSI-Evaluation on the main parallel Persian-English corpus. The CSI

[2] Persian_poems_corpus: https://github.com/amnghd/Persian_poems_corpus/tree/master

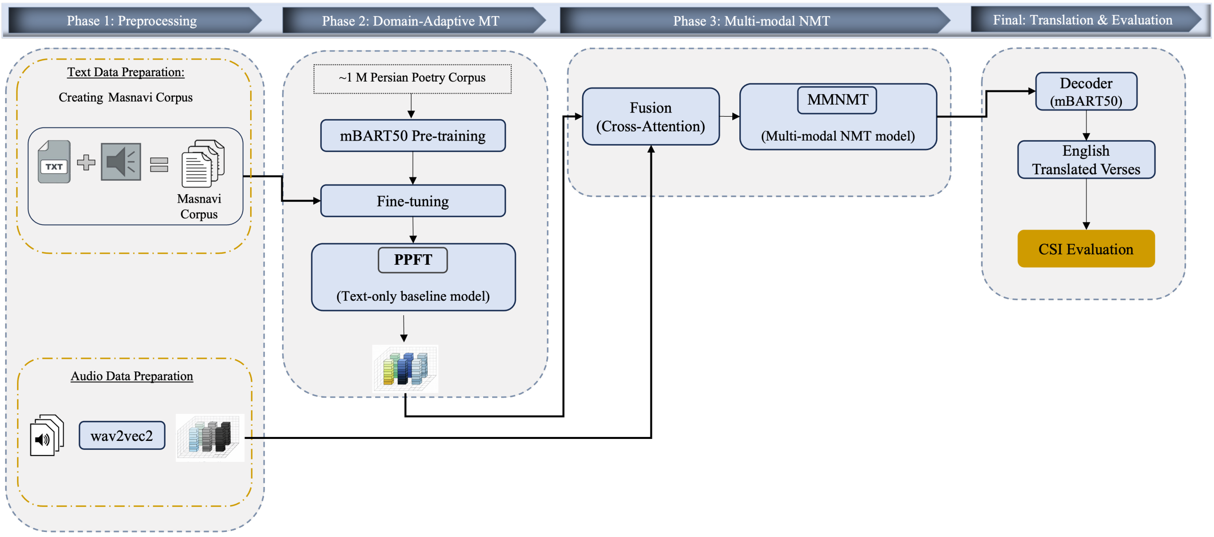


Figure 2: Overview of our proposed multi-modal translation system developed for Persian poetry translation to English.

Modality	Model	BLEU	ChrF++	BERTScore	COMET
Txt	PPFT	9.85	32.77	0.876	0.579
Txt + Audio	PPFT + Wav2Vec2	17.95	42.95	0.894	0.635

Table 2: Evaluation results for text-only baseline (PPFT) and multi-modal NMT (PPFT + Wav2Vec2) on Persian–English translation of the *Masnavi* corpus.

evaluation is done on the both uni-modal and multi-modal scenarios. Across all automatic metrics, the MMNMT demonstrates substantial and consistent improvements compare to the text-only model as elaborated in Table 2. Although the PPFT model performance metrics in Table 2 are reasonable for a low-resource poetic MT baseline, they are clearly limited. BLEU is conservative (paraphrase sensitivity), yet COMET and ChrF++ confirm the model is still missing many correct realizations.

By integration of the audio recitations and building up the proposed MMNMT system, the BLEU score and the the ChrF++ metric increased significantly. Such improvements indicate an enhanced management of morphological variations and character-level accuracy. Such character-level precision holds particular importance for Persian-to-English translation tasks. The semantic metrics like BERTscore and COMET are also presented, showing an increase from 0.876 to 0.894, and from 0.579 to 0.635, respectively. These two metrics are less sensitive to exact lexical overlap and more aligned with meaning and semantic preservation. This clearly indicates that our proposed MMNMT system produces translations that are semantically closer to the references. Furthermore, the consistent gains across all presented metrics, indicate that incorporating audio information enhanced both the fluency and the semantic adequacy of translations.

It is noteworthy that although the results have improved, the overall scores are still moderated compared to high-source prose translation tasks. This is because most automatic metrics can not fully measure the extent in which cultural and poetic meanings are preserved. For example, the conservative nature of metrics like BLEU often fails to capture the semantic and artistic depth of poetic translations. As noted by Cadotte et al. (2024), quantitative scores may not reflect the "actual quality" of poetic MT, as translations can be correct and culturally grounded even when they differ significantly from the reference string. Persian poetry includes culture-specific terms. Translating these correctly may not raise metric scores for word overlap or meaning similarity, but it is essential for understanding and keeping the cultural value of the text. This reinforces the necessity of defining a CSI-Recall metric, as investigated in the next section. Such metric provides a diagnostic of cultural fidelity that is not captured by standard evaluation metrics.

Model	Lexicon Scope	Coverage(%)	CSI-Recall (%)
PPFT	strict_core_top	24.16	61.60
PPFT + Wav2Vec2	strict_core_top	24.16	82.04
PPFT	soft_core_top	24.16	66.94
PPFT + Wav2Vec2	soft_core_top	24.16	89.04
PPFT	broad_rank	69.66	51.77
PPFT + Wav2Vec2	broad_rank	69.66	75.73

Table 3: CSI-aware evaluation of text-only baseline (PPFT) and multi-modal NMT (PPFT + Wav2Vec2) on the *Masnavi* corpus across three lexicon versions: strict_core (high-confidence gold-standard spans), soft_core (relaxer version with all tagged spans and strong alignment), and broad (maximum coverage with minimal confidence threshold).

6.1 CSI-Aware Evaluation

Table 3 presents the results of a culture-aware evaluation using our CSI taxonomy. This evaluation verifies whether culturally important Persian terms are correctly translated. Both the coverage and CSI-Recall metrics are reported. Coverage shows how many CSI words our dictionary knows. This coverage has nothing to do with our models, it only depends on the lexicon. Therefore, based on only what the lexicon can judge, the coverage of 24.16% seems low. This means that only 24% of CSI spans are lexicon-covered. Recall on the other hand means that out of the CSI words we can check, how many did the model translate correctly.

In all versions of our lexicons, the proposed MM-NMT (e.g., PPFT + Wav2Vec2 model) achieved higher recall than the uni-model PPFT. For instance, as it is shown in Table 3, in the strict core setting, CSI-Recall goes up from 61.60% to 82.04%. Similar improvements appear in the soft core and broad lexicon settings, where recall rises from 66.94% to 89.04%, and from 51.77% to 75.73%, respectively. These results show that the multi-modal model maintains considerably more cultural information when exact cultural rendering is required.

These gains illustrate that audio information helps guide the model to a more culturally appropriate translation. As discussed earlier, the act of recitation plays a vital role in how meaning is conveyed through rhythm, tone, pauses, and emphasis, elements that don't appear in writing, specifically for poetry that are hard to read and understand like *Masnavi* of Rumi. These sound-based features reveal much about the poem's structure, mood, and the poet's intentions. The written Persian language, particularly in classical poetry, is naturally ambiguous.

When audio features are integrated with text, translation models gain access to expressive and structural details that written words alone cannot capture. As supported by the results provided in both Table 2, and Table 3, the resultant multi-modal model enables translations that better preserve the depth, and intent of Persian poetry, especially in rich, culturally layered works from the classical tradition.

It is important to note two limitations of this evaluation. First, the core coverage is only 24%, which means the metric currently judges limited slice of CSI spans. Second, the CSI metric is currently recall-only over lexicon-covered spans (no precision), so it does not penalize hallucinated CSI words or incorrect usage. In order to validate that these gains are not caused by insignificant effects, we also demonstrated COMET/BERTScore improvements presented in Table 2. Based on the results, switching from text-only baseline to a multi-modal model that used the audio recitation has achieved a significant gain with a large and consistent improvement across all metrics.

7 Future Work

This work focused solely on using CSI lexicon as an evaluation step. In future work, we would like to investigate the impact of using our refined lexicon as an analytical resource for examining cross-cultural transfer phenomena in our proposed models. In addition, further refinement of the created CSI lexicon, particularly through expanded coverage, would strengthen its reliability. Finally, because automatic metrics such as BLEU, ChrF++, and BERTScore do not fully capture poetic fidelity, we will incorporate structured human evaluation protocols.

Conclusion

This study explores both the computational and cultural challenges of translating Rumi's *Masnavi-ye-Ma'navi*, a Persian mystical poetry known for its intentional ambiguity and rich symbolism. Since NMT systems show inadequacy in the low-resource poetic settings, we developed the first multi-modal translation framework specifically for classical Persian poetry. Our main contributions are: (1) creating a new parallel corpus with 26,571 pairs of Persian-English verse from all six books of the *Masnavi*, each accompanied by aligned audio recitations; (2) establishing a domain-adapted baseline through pre-training on Persian poetry corpus followed by fine-tuning on our *Masnavi* dataset (train set); (3) improving the model by integrating acoustic features via cross-attention fusion mechanisms; and (4) creating a CSI taxonomy and lexicon for the evaluation of cultural preservation in translation outputs.

The experiments show that adding audio greatly improves the results across multiple dimensions. Standard metrics show notable gains, BLEU increased from 9.85 to 17.95, while ChrF++ goes up from 32.77 to 42.95. More significantly, our CSI-aware evaluation demonstrates that multi-modal models better preserve cultural fidelity, with recall rising from 61.60% to 82.04% under strict lexicon constraints. These findings confirm that audio features and information help the model understand meaning more accurately and maintain cultural authenticity. This work establishes a foundation for culturally-aware machine translation of classical poetry. It demonstrates that multi-modal approaches can preserve the literary richness and mystical depth that are central to Persian poetry. Future research focus on expanding lexicon coverage, investigating audio's contribution through systematic ablation studies, and using the lexicon as an extra training resource to strengthen cultural understanding in translation models.

References

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, and 1 others. 2021. XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Antoine Cadotte, Nathalie André, and Fatiha Sadat. 2024. Machine translation through cultural texts: Can verses and prose help low-resource indigenous models? In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 121–127, Bangkok, Thailand. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, and Smaranda Muresan. 2021. Don't go far off: An empirical study on neural poetry translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7253–7265, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ruiyao Gao, Yumeng Lin, Nan Zhao, and Zhenguang G Cai. 2024. Machine translation of Chinese classical poetry: a comparison among ChatGPT, google translate, and deepl translator. *Humanities and Social Sciences Communications*, 11(1):1–10.

Mozhgan Ghassemiazghandi. 2023. Machine translation of selected ghazals of Hafiz from Persian into English. *AWEJ for Translation & Literary Studies*, 7(1).

Mojdeh Javadi, Mani Zarinebaf-Shahr, Alan Lewis, and Maryam Lewis. 2015. Masnavi.net: A full-text website of the masnavi in persian, english, and turkish. http://masnavi.net/. Accessed: 2026-01-11.

Reza Khanmohammadi, Mitra Sadat Mirshafiee, Yazdan Rezaee Jouryabi, and Seyed Abolghasem Mirroshandel. 2023. Prose2poem: The blessing of transformers in translating prose to persian poetry. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(6).

Alan Lewis. 2015. The masnavi. https://www.dar-al-masnavi.org/masnavi.html. Accessed: 2026-01-11.

Marian Lupascu, Ana-Cristina Rogoz, Mihai Sorin Stupariu, and Radu Tudor Ionescu. 2025. Large multimodal models for low-resource languages: a survey. *arXiv preprint arXiv:2502.05568*.

Reynold A. Nicholson and Sufism.org. 2013. Rumi: Masnavi, book i (version 1.0). https://sufism.org/wp-content/uploads/2013/12/Rumi-Book-I-Version-1.0.pdf. Accessed: 2026-01-11.

Shantipriya Parida, Ondřej Bojar, Idris Abdulmumin, Shamsuddeen Hassan Muhammad, and Ibrahim Said Ahmad. 2024. Findings of WMT2024 English-to-low resource multimodal translation task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 677–683, Miami, Florida, USA. Association for Computational Linguistics.

Abdorreza Rafiee. 1991. *Variables of communicative incompetence in the performance of Iranian learners of English and English learners of Persian.* Ph.D. thesis, School of Oriental and African Studies (University of London).

Pawan Rajpoot, Nagaraj Bhat, and Ashish Shrivastava. 2024. Multimodal machine translation for low-resource Indic languages: A chain-of-thought approach using large language models. In *Proceedings of the Ninth Conference on Machine Translation*, pages 833–838, Miami, Florida, USA. Association for Computational Linguistics.

Nikta Gohari Sadr, Sahar Heidariasl, Karine Megerdoomian, Laleh Seyyed-Kalantari, and Ali Emami. 2025. We politely insist: Your LLM must learn the Persian art of Taarof. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1819–1838.

Amir Artaban Sedaghat. 2020. Translating Rumi through the prism of ideology. *Iran Namag*, 5(2):4–34.

Mohammadreza Shahrestani and Mostafa Haghir Chehreghani. 2025. Prosody recognition in persian poetry. *Speech Communication*, 170:103222.

Stefanie Stadler. 2012. Cross-cultural pragmatics. *The encyclopedia of applied linguistics*, pages 1–8.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Madhavendra Thakur. 2025. Culturally-grounded chain-of-thought (CG-CoT): Enhancing LLM performance on culturally-specific tasks in low-resource languages. *arXiv preprint arXiv:2506.01190*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Shanshan Wang, Derek Wong, Jingming Yao, and Lidia Chao. 2024. What is the best way for ChatGPT to translate poetry? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14025–14043, Bangkok, Thailand. Association for Computational Linguistics.

Bajian Xiang, Shuaijiang Zhao, Tingwei Guo, and Wei Zou. 2025. Understanding the modality gap: An empirical study on the speech-text alignment mechanism of large speech language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5187–5202.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking machine translation with cultural awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13078–13096, Miami, Florida, USA. Association for Computational Linguistics.