Exploration of adaptation of XAI methods for Automatic Speech Recognition

Arunima Chaurasia

***Abstract*—This research explores visual domain explainability methods for extending explainability to Automatic Speech Recognition (ASR) systems. We address this gap by adapting AtMan's [1] attention manipulation approach to analyze how ASR systems, specifically Whisper [2], process audio inputs and generate transcriptions.**

Our methodology involves systematically perturbing attention scores in Whisper's encoder layers and measuring the resulting changes in cross-entropy loss at both token and sentence levels. By applying different suppression factors and window sizes to attention scores along the temporal axis, we generate local explanation maps that reveal the relationship between temporal frames of log-mel spectrograms and transcribed tokens. We evaluate our approach on the LibriSpeech [3] dataset using Whisper's 'tiny' variant.

The results demonstrate that frames containing speech content show significantly higher influence compared to silent segments, and the model's token predictions are most heavily influenced by immediately preceding temporal frames. Additionally, we observe that larger suppression windows and more aggressive suppression factors lead to greater disruption in model predictions. While our adaptation successfully provides temporal explanations, it is limited by its inability to provide frequency-wise interpretations due to Whisper's architecture.

This work demonstrates that AtMan can be adapted to ASR models and provides an efficient method for understanding temporal dependencies in transformer-based speech recognition models, while also highlighting the challenges and limitations in adapting the proposed method to audio processing tasks.

***Index Terms*—Explainable AI, Automatic Speech Recognition, Transformer-based models.**

I. Introduction

The set of methodologies and techniques when applied to AI systems, helping in interpretations of the decisions made by the system are called Explainable Artificial Intelligence (XAI) methods. With the increase in use of deep neural network based Automatic Speech Recognition (ASR) systems in our day-to-day life, the need has arisen for a certain amount of trustworthiness for the user to accept the results generated by the system [4]. Current ASR systems often operate as "black boxes" with limited insights into their decision-making processes. The state of the art ASR systems are transformer-based models that are increasingly outperforming ASR systems based on classical approaches like Hidden Markov Models [5]. Most of the research on explainability has been concentrated on visual and structured data. There is a lack of research for Explainable AI specifically in ASR [4]–[6]. This research addresses the mentioned gap by exploring adaptibility of visual domain interpretability methods for interpretability of transformer-based ASR models.

For the ASR model, we focus on transfomer-based end-to-end ASR models, specifically the Whisper [2] model. End-to-end models use a single neural network to map audio inputs directly to text outputs. These models typically use deep learning architectures that jointly learn acoustic, pronunciation, and language representations.

In this research, we adapt one such method AtMan [1] as our explainability method to adapt for the Whisper model [2]. Originally developed for vision transformers and natural language processing tasks, AtMan (Attention Manipulation) is a novel approach that leverages attention manipulation techniques to provide insights into transformer model decisions. It works by systematically introducing perturbations through attention manipulations across different transformer layers during forward propagation, helping to reveal the model's internal decision-making process. The method was initially demonstrated on transformer based language models for image classification through prompt and text generation tasks, where it successfully identified relevancy of tokens for model predictions. We adapt the method for end-to-end ASR Whisper model and evaluate it on Librispeech [3] dataset, examining the relevance of particular time sequence for the output tokens generated by the model.

A. Motivation

The motivation for this research is to provide a method to interpret the decisions made by transformer-based ASR models. This is crucial for several reasons. First, as ASR systems become increasingly integrated into critical applications like healthcare, legal transcription, and emergency response systems [4], understanding their decision-making process becomes vital for ensuring reliability and accountability. Second, in cases where ASR systems make errors, having interpretable models allows developers to identify the root causes of these mistakes, whether they stem from acoustic ambiguities, background noise, or linguistic complexities. Third, explainability builds trust with end-users by providing transparency about how their speech is being processed and transcribed, particularly important in privacy-sensitive contexts. Additionally, for developers and researchers, understanding the internal workings of these models is essential for systematic improvement and debugging of ASR systems [5], as well as for ensuring compliance with emerging AI regulations that may require algorithmic transparency.

*Submitted to the Department of Computer Science at Hochschule Bonn-Rhein-Sieg in partial fulfilment of the requirements for the degree of Master of Science in Autonomous Systems

†Supervised by Prof. Dr. Sebastian Houben (Hochschule Bonn-Rhein-Sieg) and Roman Bartolosch, M.Sc. (FKIE)

‡Submitted in February 2025

B. Problem Statement

This research investigates the adaptation of successful XAI methods from the visual, structured, and audio domains, which have not yet been applied to ASR systems, to improve accountability and transparency. The primary objective is to assess the feasibility of directly adapting these methods for ASR explainability. This investigation will focus on identifying the necessary modifications for direct adaptation and evaluating the relevance and effectiveness of the resulting explanations.

- Adapting visual domain interpretability methods, particularly AtMan [1], for ASR applications.
- Understanding how audio input in relation to attention mechanisms in transformer-based ASR models influence transcription decisions.

C. Proposed Approach

Upon reviewing existing methods, we identified AtMan [1] as a novel technique that examines model behavior by introducing controlled perturbations to attention weights within transformer layers. By systematically altering attention patterns and observing their effects on the model's predictions, AtMan reveals which tokens, when suppressed, have the greatest impact on subsequent tokens, or which previous tokens most significantly influence a particular token. This provides valuable insights into the model's decision-making process. We apply this method to elucidate the decisions made by the Whisper model [2] on the Librispeech [3] dataset. By perturbing the attention weights in the encoder layers and observing the resulting changes in the model's predictions, we aim to identify the audio segments that are most critical to the model's decision-making process.

II. Related Work

A. Explainability in Audio Domain

According to Akman and Schuller [4], audio explainability methods can be categorized into two groups: generic methods adapted for audio [7]–[9] and specialized methods designed specifically for audio models [5], [6], [10], [11]. In AudioLIME, Haunschmid et al. [11] adapted LIME [12] for a music tagging system. It uses source separation to provide listenable explanations, enabling users to understand the model's decisions by hearing the separated components that influenced the decision most.

In the context of speech processing, Sivasankaran et al. [13] applied DeepSHAP to speech enhancement models, revealing which frequency-time bins in the input spectrogram influence the model's masking decisions.

Frommholz et al. [14] utilized the post-hoc interpretability method LRP in conjunction with the Discrete Fourier Transform (DFT) to explore how various input representations influence the decision-making process of the model. They compared the relevance of heatmaps from models trained on different input forms to assess how these representations affect the model's output.

Recent work by Parekh et al. [15] has also explored the use of Non-negative Matrix Factorization (NMF) for interpretability in audio classification networks, providing a novel approach to understanding model decisions through decomposition of audio features.

B. Explainability in ASR

The field of explainability in ASR was pioneered by Wu, Bell and Rajan [6] with the X-ASR framework. The authors adapted generic XAI methods to ASR models, namely Statistical Fault Localization (SFL) [16], Local Interpretable Model-Agnostic Explanations (LIME) [12] and causal explanations [17]. They emphasize on the importance of qualitative assessment of the explanations, evaluated by the size and consistency of explanations of transcriptions generated across various ASR models for the respective methods. Wu, Bell and Rajan [18] uses state of the art visual domain XAI method Local Interpretable Model-Agnostic Explanations (LIME), adapted from the image classification domain to ASR model for phoneme recognition and demonstrates reliable explanations 96% of the time in its top three audio-segments in a controlled evaluation setting. A significant advancement came with GRAD-SAS by Sun et al. [5], which adapted Grad-CAM [8] for speech recognition transformers. This technique involves calculating gradients by deriving classification scores and combining attention matrices with their gradients through inner-product and summation operations to visualize the weighted attention on input tokens, providing insights into the model's focus during transcription.

Recent research by Javadi et al. [19] has focused on understanding word-level relationships in ASR outputs. They proposed a novel approach for word-level error ASR quality estimation by analyzing attention patterns in reference-free metrics. This method not only helps in understanding the relationship between input audio and transcribed words but also enables efficient corpus sampling and post-editing. The approach demonstrates how attention based error analysis can be leveraged to provide explanations at the word level.

C. Explainability in Audio Transformers

Explainability in Audio Transformers is crucial as they outperform other deep neural networks for audio-specific tasks as shown by Prabhavalkar et al. [20]. AttHear by Akman and Schuller [10] explains audio transformers by generating listenable explanations using inverse Short-time Fourier Transform (STFT). It pre-processes the audio to extract features, uses a pre-trained transformer to compute attention scores, and applies Non-negative Matrix Factorization (NMF) weighted by these scores to identify the important audio parts. Bicer et al. [21] enhanced an acoustic scene classification model's interpretability using Grad-CAM and guided backpropagation, providing visual insights into which parts of the input contributed most to the model's decisions. As for the application of explainability in Audio domain, Vitale et al. [22] utilizes explainability techniques to analyze emergent syllables in end-to-end ASR systems. They used model explainability techniques to investigate how these recognizers identify and process syllables, providing insights into the model's internal mechanisms and improving our understanding of how speech patterns are learned and recognized by these advanced models. Research

by Singla et al. [23] has also explored probing transformer representations for language delivery and structure, providing insights into how these models process and understand speech patterns. This work complements traditional explainability approaches by revealing the internal representations learned by transformer models and their relationship to linguistic features.

D. Explainability in Visual Transformers

Our research focuses on adapting visual domain methods for ASR, particularly using end-to-end transformer models. Therefore, understanding methods in visual transformers is essential. These methods are categorized into feature-attribution, attention-based, pruning-based, inherently explainable methods, and other tasks as described by Kashefi et al. [24]. Feature-attribution techniques like Grad-CAM [8] and Layer-wise Relevance Propagation (LRP) [7] are most commonly utilized methods in the domain. Grad-CAM generates heatmaps by utilizing model gradients to highlight class-relevant areas in input images, thereby improving model transparency. The approach by Chefer et al. Chefer, Gur and Wolf [25] offers a generic attention-model explainability method for bimodal and encoder-decoder transformers, focusing on how input components influence model decisions through attention mechanisms. T-TAME by Ntrougkas, Gkalelis and Mezaris [26] is another attention-based method that employs unsupervised learning to train attention mechanisms using feature maps, facilitating the interpretation of model processing and feature prioritization. R-Cut by Niu et al. Niu et al. [27] creates alternative activation maps from normalization layers, perturbs input images, and uses similarity scores from a pre-trained model to weigh these maps. It uses class-aware patch tokens to form a weighted graph, where nodes represent semantic features, and generates explainability maps by partitioning with graph eigenvectors, enhancing class-specific decision understanding. LeGrad by Bousselham et al. [28] utilizes the self-attention mechanism in ViT [29] by computing gradients concerning attention at each layer, summing up explainability maps from each layer to produce a final map. Ali et al. [30] demonstrate that traditional XAI methods often produce unreliable results when applied to transformer models, and propose an extension to Layer-wise Relevance Propagation (LRP) [7]. Their enhanced method introduces conservative propagation rules for transformer architectures, resulting in more consistent and reliable explanations across different model components. For transformers, most methods leverage attention mechanisms. AtMan by Deiseroth et al. [1] proposes understanding transformer predictions through memory-efficient attention manipulation. It is essentially developed for generative transformers that explore how different parts of the input affect the model's predictions. It manipulates attention mechanisms by perturbing token embeddings and measures changes in model output. By comparing these perturbed embeddings using cosine similarity, AtMan [1] identifies which input tokens are crucial for the model's decisions without relying on traditional backpropagation methods.

Our research enhances the existing body of work on explainability for ASR and audio transformers by adapting and expanding the AtMan method by Deiseroth et al. [1], which was initially created for visual transformers, to the ASR domain. Previous studies such as X-ASR by Wu, Bell and Rajan et al. [6] and GRAD-SAS by Sun et al. [5] have established the groundwork for ASR explainability. However, our method emphasizes manipulating attention in the embedding space to gain deeper insights into transformer-based ASR models, while avoiding the computationally intensive backpropagation typically used in attention-based explainability techniques.

Unlike traditional methods that rely on gradient-based explanations or post-hoc analysis, our adaptation of AtMan systematically perturbs attention weights within the transformer layers to reveal the internal decision-making process of the model. This approach allows us to identify the most critical audio segments influencing the model's predictions, thereby enhancing the interpretability of end-to-end ASR systems.

III. Background

A. Explainable AI

Explainable AI (XAI) is a field of AI research that focuses on making AI systems more understandable and interpretable. It aims to provide explanations for AI model predictions, decisions, and behaviors, helping users understand how the model arrives at its conclusions. It can be grouped into three categories: post-hoc vs intrinsic, local vs global, model agnostic vs model specific explanations. Post-hoc explanations are generated after the model has been trained, while intrinsic explanations are built into the model architecture. Local explanations focus on explaining the model's decision for a specific input, while global explanations focus on explaining the model's behavior over a range of inputs. Model agnostic explanations are independent of the model architecture, while model specific explanations are specific to the model architecture [31].

B. Audio Waveform

An audio waveform represents the temporal evolution of sound pressure variations as a digital signal. It plots amplitude against time, where amplitude corresponds to the instantaneous sound pressure level. In digital form, the waveform consists of discrete samples captured at a fixed sampling rate, known as sampling. Sampling is the process of converting a continuous analog signal into a sequence of discrete numerical values by measuring the signal's amplitude at regular time intervals [32]. The sampling rate, measured in Hertz (Hz), determines how many samples are taken per second - for example, CD-quality audio uses 44.1 kHz sampling rate, meaning 44,100 samples are taken every second. While waveforms provide a complete representation of the audio signal, Automatic Speech Recognition (ASR) systems typically transform them into more compact features like spectrograms for efficient processing [33].

C. Log-Mel Spectrograms

A spectrogram is a visual representation of the spectrum of frequencies in a sound signal as they vary with time. It is

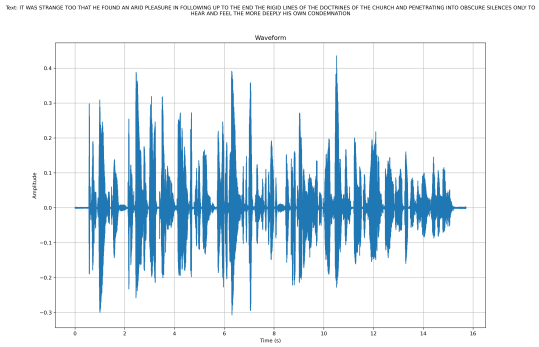

Fig. 1: Audio waveform of an example from LibriSpeech dataset

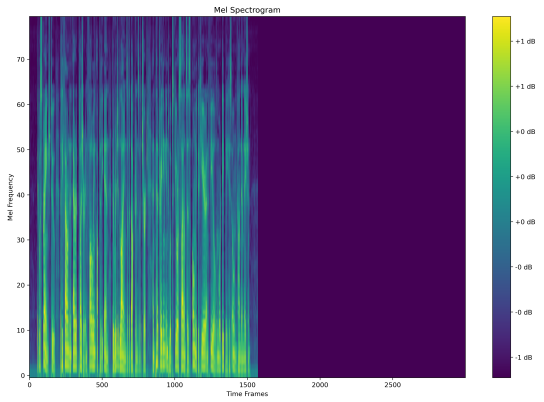

Fig. 2: Log-Mel spectrogram with 80 mel bins

created by applying a Short-Time Fourier Transform (STFT) to the audio signal, which breaks it down into its frequency components over short time windows.

A Mel spectrogram is a type of spectrogram where the frequency axis is converted to the Mel scale, a perceptual scale of pitches designed to approximate the way humans perceive sound frequencies [34]. It is based on the idea that humans do not perceive pitch linearly; instead, our perception of pitch is more closely related to a logarithmic scale. This transformation emphasizes frequencies that are more relevant to human perception, making it particularly useful for audio processing tasks. The Mel scale is approximately linear below 1 kHz and logarithmic above 1 kHz [35]. The Mel filter bank consists of a series of overlapping triangular filters, each corresponding to a "Mel bin." Each Mel bin represents a range of frequencies, and the output of each filter is the sum of the energy in that frequency range. The number of Mel bins determines the resolution of the Mel spectrogram.

The need for Log-Mel spectrograms arises from the fact that the human ear perceives loudness on a logarithmic scale. By taking the logarithm of the Mel spectrogram, we obtain a Log-Mel spectrogram, which better aligns with human auditory perception. This representation is particularly useful for ASR models as it captures the essential features of speech while reducing dimensionality. A Log-Mel spectrogram can be interpreted by examining its three main dimensions:

- The horizontal axis represents time progression
- The vertical axis shows frequency in Mel scale (emphasizing lower frequencies that are more relevant to human speech)
- Color intensity indicates the energy/amplitude at each time-frequency point, with brighter colors representing higher energy

In log-mel spectrograms, silent periods are visible as darker regions with minimal energy across frequencies, reflecting the absence of vocal activity. This visual pattern in spectrograms helps in distinguishing different phonetic elements and analyzing speech characteristics.

D. Attention Mechanisms

Attention mechanisms are integral to transformer models, a neural network architecture highly effective for sequence-to-sequence tasks such as Automatic Speech Recognition (ASR), as initially proposed by Vaswani et al. [36]. These mechanisms enable models to dynamically prioritize different segments of the input sequence, assigning varying importance to each token. In transformers, attention mechanisms facilitate the capture of long-range dependencies and contextual information, essential for interpreting complex input sequences like speech [36].

Each input token is transformed into three vectors: the query Q, the key K, and the value V. The query vector represents what information the token is looking for, the key vector represents what information the token contains that others might want to query, and the value vector contains the actual content that will be aggregated in the attention computation. These vectors are obtained by multiplying the input token embedding with learned weight matrices - trainable parameters that are optimized during model training to capture meaningful transformations of the input embeddings. The weight matrices start with random values and are gradually adjusted through backpropagation to learn optimal transformations for the specific task.

Operation:

- **Self-Attention Mechanism:** In self-attention, the Q, K, and V matrices are all derived from the same input sequence. This allows the model to focus on different parts of the sequence itself, capturing dependencies and relationships within the sequence.
- **Cross-Attention Mechanism:** In cross-attention, the Q matrix comes from one sequence (e.g., the decoder in a transformer), while the K and V matrices come from another sequence (e.g., the encoder). This allows the model to focus on relevant parts of a different sequence, facilitating tasks like translation or multi-modal processing.
- **Attention Scores:** The attention scores are computed by taking the dot product of the query and key vectors, followed by a softmax function to normalize the scores. These scores determine how much attention each part of the input sequence receives.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where d_k is the dimension of the key vectors. The scaling factor $\frac{1}{\sqrt{d_k}}$ prevents the dot products from growing too large in magnitude, which could push the softmax function into regions with extremely small gradients.

- **Output:** The final output of the attention mechanism is a weighted sum of the value vectors, where the weights are the attention scores. This output is then used in subsequent layers of the model to make predictions or generate outputs.

E. Perturbations

Perturbations involve introducing small changes to the input data or model parameters to analyze their impact on the model's predictions. This technique is used in explainability methods to identify which parts of the input are most influential in the decision-making process, providing insights into the model's internal workings [31].

F. Cross Entropy Loss

Cross entropy loss is a common loss function used in classification tasks, including ASR. It measures the difference between the predicted probability distribution and the true distribution, penalizing incorrect predictions. Minimizing cross entropy loss helps improve the accuracy of the model's predictions [37], [38].

The formula for cross entropy loss is:

$$L(y, \hat{y}) = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \tag{2}$$

where y is the true distribution (often represented as a one-hot encoded vector), $\hat{y}$ is the predicted probability distribution, and N is the number of classes.

G. Logits

In transformer models, logits are the raw output scores produced by the final linear layer before any normalization is applied. For ASR transformers particularly Whisper, these logits represent the model's pre-softmax predictions for each possible token (typically characters or subword units) in the vocabulary at each timestep. The logit values can be positive or negative and their magnitude indicates the model's confidence - larger values suggest stronger predictions for particular tokens.

The logits are computed through a linear transformation of the final hidden states:

$$\text{logits} = W_{\text{vocab}} h + b \tag{3}$$

where h is the hidden state vector from the final transformer layer, W_{vocab} is a learned weight matrix that maps the hidden state dimension to the vocabulary size dimension (i.e., $W_{\text{vocab}} \in \mathbb{R}^{d_{\text{vocab}} \times d_{\text{model}}}$), and b is a bias term, so logits are the size of the vocabulary for each forward pass of decoder while predicting the next token.

These raw logits are then transformed into probabilities using the softmax function:

$$p(y_i) = \frac{\exp(\text{logit}_i)}{\sum_j \exp(\text{logit}_j)} \tag{4}$$

where $p(y_i)$ is the probability of token i. This normalization ensures the probabilities sum to 1 while preserving the relative rankings of the logits [39]. In ASR systems, these probabilities guide the decoder in selecting the most likely transcription sequence [2], and the logits are used to compute the cross entropy loss.

H. AtMan

As we explore multiple visual domain methods for explaining transformer based ASR models, we found AtMan by Deiseroth et al. [1] to be a promising method for our task, since it was modality agnostic. It is a local explanation method that operates by perturbing the transformer's attention mechanism rather than the input space directly. Its core principle is to manipulate attention scores during the forward pass and observe how these manipulations affect the model's output logits at the decoder. Specifically, it suppresses attention to specific tokens (or groups of correlated tokens) and measures the resulting changes in the cross-entropy loss for target predictions. The magnitude of change in the output indicates the importance of the perturbed tokens for the specific prediction being explained. This approach is more efficient than traditional perturbation methods because it works in embedded token space rather than input space, and more memory-efficient than gradient-based methods since it only requires forward passes through the model instead of backpropagation. The method quantifies the influence of input tokens by comparing the model's original prediction with predictions under attention perturbation, effectively creating relevance maps that explain which parts of the input were most important for specific outputs.

IV. Methodology

In our adaptation of AtMan for Whisper, we adapt the proposed method of single-token attention manipulation from natural language processing to the audio domain, specifically focusing on Whisper's log-mel spectrogram input. While AtMan also proposes a method for handling correlated tokens in visual data, we do not apply this aspect to audio processing for two key reasons:

1) Unlike 2D images that are divided into patches, audio input does not have an inherent patch-based structure. In case of Whisper, it processes chunks of 30s audio at a time. We assume that multiple chunks of audio are not correlated to each other.
2) Although the audio is converted into a 2D log-mel spectrogram representation, the Whisper encoder processes this representation holistically rather than as discrete patches for generating the embeddings.

In our adaptation of single-token attention manipulation for the audio domain, we modify the attention scores in the attention layers of the Whisper model's encoder by applying suppression factors (values between 0 and 1). These suppression factors indicate the degree of attention suppression, where values closer to 0 indicate stronger suppression and 1 indicates no suppression. For each layer and attention head u, we modify the pre-softmax attention scores as:

$$\tilde{H}_{u,*,*} = H_{u,*,*} \circ (f_i) \tag{5}$$

where:

- H represents attention scores
- f_i is suppression factor for token i
- $\circ$ is Hadamard (element-wise) product
- u indexes over all layers and attention heads

After applying attention manipulation, we analyze its impact by measuring the model's cross-entropy loss at both token and sentence levels to quantify the influence of suppressed frames. The audio input is represented as a sequence of log-mel spectrogram frames $x = [x_1, ..., x_N]$, where each x_i represents a temporal frame and N is the total number of frames. Following the original ATMAN paper, we perform perturbations in the embedded token space where the log-mel spectrogram frames are converted to audio features. Our analysis requires maximum 1500 forward passes through Whisper's 'tiny' model, corresponding to the audio context of the encoder model and the number of frames after the encoder's convolutional layers downsample the original 3000 log-mel spectrogram input frames. Through systematic frame-by-frame perturbation and loss measurement, we can pinpoint the temporal segments that have the strongest impact on the final transcription.

In our implementation, rather than suppressing attention scores for individual frames, we apply suppression across a window of consecutive frames. This windowed approach means that for each perturbation instance, we simultaneously suppress attention scores for multiple adjacent frames within the specified range. The cross-entropy loss is computed for each perturbation by evaluating the difference between the model's predicted logits and the target transcription. This evaluation happens at both the token and sentence level, providing insight into the perturbation's effects at different scales of granularity. The initial computation yields loss values for 1500 frames, which corresponds to the downsampled sequence length after the encoder's convolutional layers. To align with the original 3000-frame temporal sequence of the audio input, we upsample the loss values through linear interpolation, effectively doubling each frame to match the original temporal resolution. This upsampling reverses the stride-2 downsampling performed by the encoder's convolutional layers. The mathematical formulation of the cross-entropy loss, adapted from the original ATMAN paper, is expressed as:

$$\begin{aligned} L_{target}(x, \theta) = \sum_t L_{target}(x_{<t}, \theta) = \\ \sum_t -\log \text{softmax}(h_\theta(x_{<t}) W_\theta^T)_{target_t} \end{aligned} \tag{6}$$

where:

- L_{target} is the cross-entropy loss for a specific target transcription
- x represents the input log-mel spectrogram frames
- θ represents the model parameters
- t is the position in the output sequence
- $x_{<t}$ represents input frames up to position t
- h_θ denotes the Whisper model (encoder-decoder)
- W_θ is the learned embedding matrix
- $target_t$ is the vocabulary index of the t-th target token in the transcription

For a given window of frames that we suppress, we track two types of loss that are formulated as follows:

1) Token-level loss for each token t in the transcription:

$$L^t_{token}(x, \theta^{-i}) = -\log \text{softmax}(h_\theta(x_{<t}) W_\theta^T)_{target_t} \tag{7}$$

2) Sentence-level loss aggregated over all tokens:

$$L_{sentence}(x, \theta^{-i}) = \sum_{t=1}^{T} L^t_{token}(x, \theta^{-i}) \tag{8}$$

These influence or loss values can be visualized through explanation maps at both token and sentence levels after upsampling to match the original temporal resolution, which are formulated as follows:

1) Token-level explanation maps: For each token t in the transcription, we generate a separate plot showing the influence across all frames:

$$\begin{aligned} E_{token}(x, target_t) = U((I^t_{token}(x_1, x), \\ ..., I^t_{token}(x_N, x))) \end{aligned} \tag{9}$$

where U represents the upsampling operation from 1500 to 3000 frames. This results in multiple plots, one per transcribed token, where each plot shows which audio frames were most influential for that specific token at the original temporal resolution.

2) Sentence-level explanation map: A single plot showing the aggregated influence across all frames for the entire transcription:

$$E_s(x, target) = U((I_s(x_1, x), ..., I_s(x_N, x))) \tag{10}$$

where:

- s is the sentence level
- U is the upsampling operation from 1500 to 3000 frames

These maps provide both a comprehensive view of how individual audio frames influence the complete transcription at the original temporal resolution, as well as detailed token-level insights showing the relationship between specific tokens and different corresponding audio segments.

We apply suppression only to the encoder layers for two key reasons. First, the encoder layers are specifically responsible for processing the audio features that are then passed to the decoder for generating the transcription. Second, our primary goal was to understand the relationship between audio time frames and the resulting transcription, rather than analyzing token-to-token relationships within the decoder (as is common in NLP transformer analysis that requires decoder layer suppression).

V. EVALUATION

Our experimental evaluation utilized the 'tiny' variant of Whisper [2] and the LibriSpeech [3] dataset. We computed influence functions at both the individual token and full sentence levels. Using sampled audio clips from LibriSpeech, we explored multiple parameter combinations: four suppression factors (0.3, 0.5, 0.7, and 0.9) and four window sizes for suppression (3, 5, 7, and 10 frames).

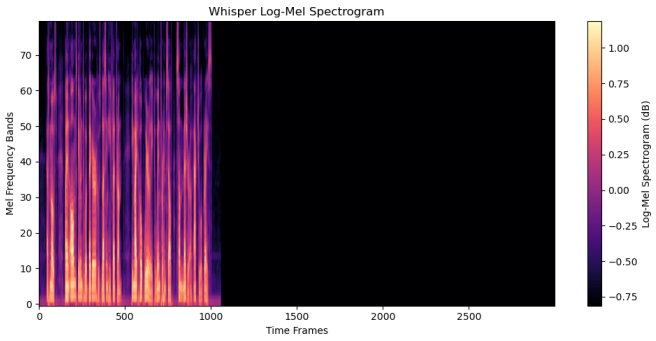

Fig. 3: Log Mel spectrogram of the audio clip for the sentence: *At most by an alms given to a beggar whose blessing he fled from he might hope wearily to win for himself some measure of actual grace*

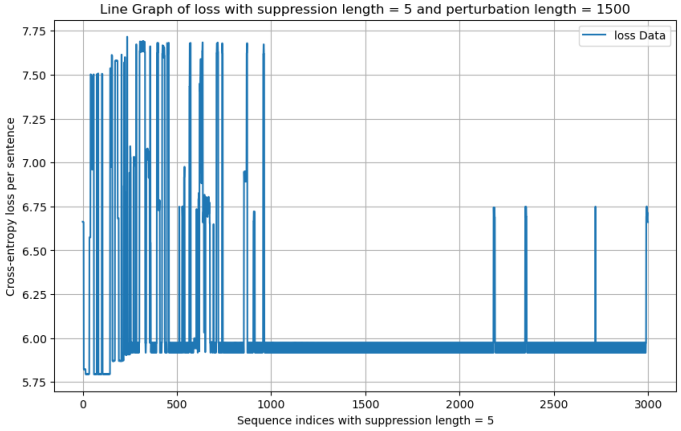

Fig. 4: Sentence-level influence map for the transcription of the audio clip in figure 3

Hypothesis 1: *Frames containing actual speech content show significantly higher influence compared to silent or padded segments of the audio.*

As shown in figure 4, we can confirm the hypothesis that frames containing actual speech content show significantly higher influence compared to silent or padded segments of the audio. The other examples also confirm this hypothesis and have been added to the appendix .

Hypothesis 2: *The temporal frames immediately preceding a token has the strongest influence on the model's prediction of that token.*

Analysis of figure 5 reveals that the model's token predictions were most heavily influenced by the temporal frames that directly preceded each token. Additionally, we observed that

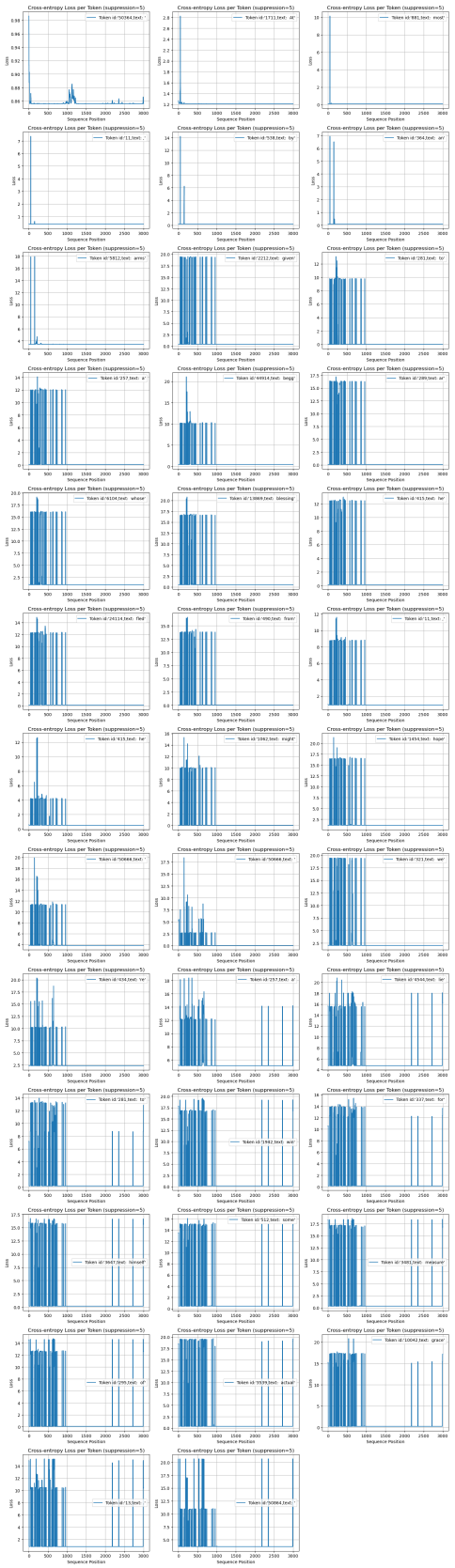

Fig. 5: Token-level influence maps for the transcription of the audio clip in figure 3

for tokens located near padded regions, the padded temporal frames also exhibited influence on the model's predictions.

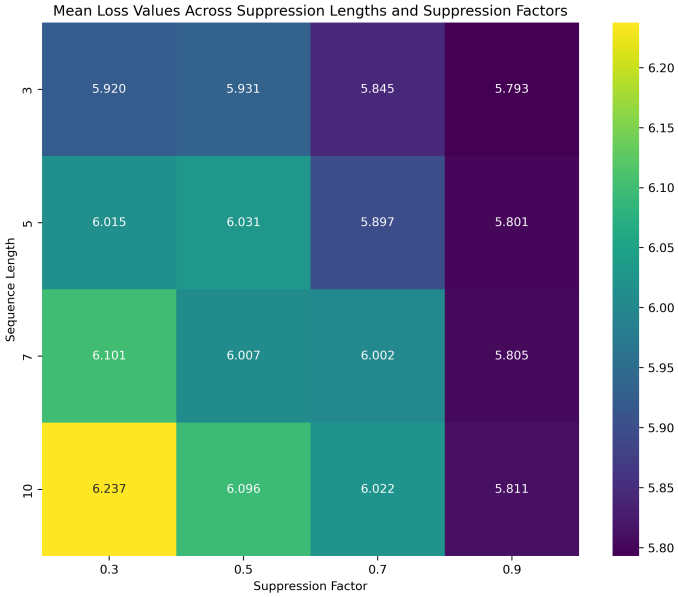

Fig. 6: Heatmap for the audio clip in figure 3 showing the mean cross entropy loss for sentence across different suppression factors and window sizes

Hypothesis 3: *Larger suppression windows and more aggressive suppression factors (values closer to zero) result in greater disruption to the model's predictions.*

Figure 6 demonstrates that both increasing the suppression window size and using more aggressive suppression factors leads to greater disruption in model predictions. This is evident from two key observations: First, when more temporal frames are suppressed through larger window sizes, the cross-entropy loss increases substantially. Second, suppression factors closer to zero apply stronger suppression to the frames, effectively diminishing their contribution to the model's attention mechanism and resulting in higher prediction errors.

In figure 7 and 8, we can analyze that even with lower suppression factors like 0.9, the cross entropy loss for larger window sizes is comparable to that of small window sizes for same suppression factor. Even with increased window size it did not impact the loss evolution much.

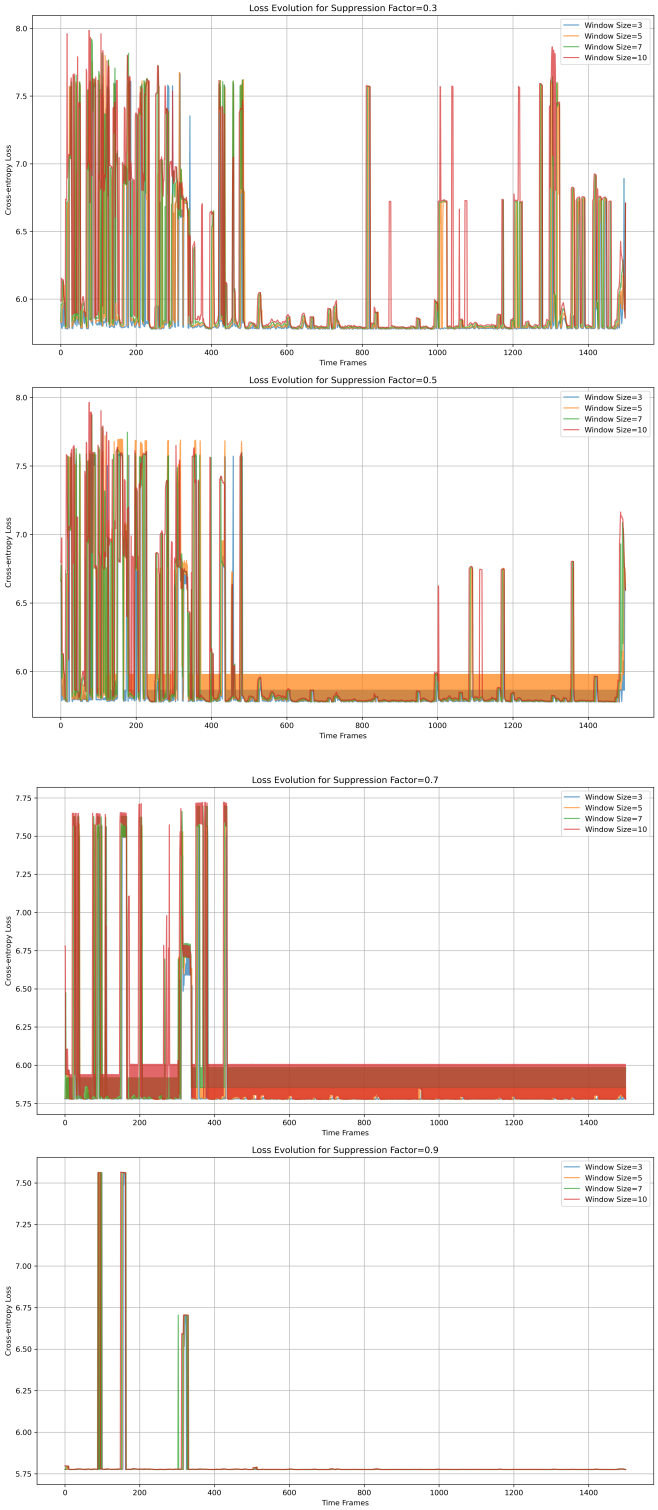

Fig. 7: Cross entropy loss for a suppression factor across different window sizes over time frames

VI. Conclusions

A. Summary

This research presents an adaptation of AtMan, an explainability framework designed for generative transformers across multiple modalities, to examine the inner workings of Whisper, a transformer architecture for automatic speech recognition (ASR). By systematically manipulating attention and introducing targeted perturbations, we investigate the model's response when attention is selectively suppressed across different temporal frames of log mel spectrogram generated from the audio input. We selected AtMan specifically because it operates through sequential forward passes, avoiding the computational

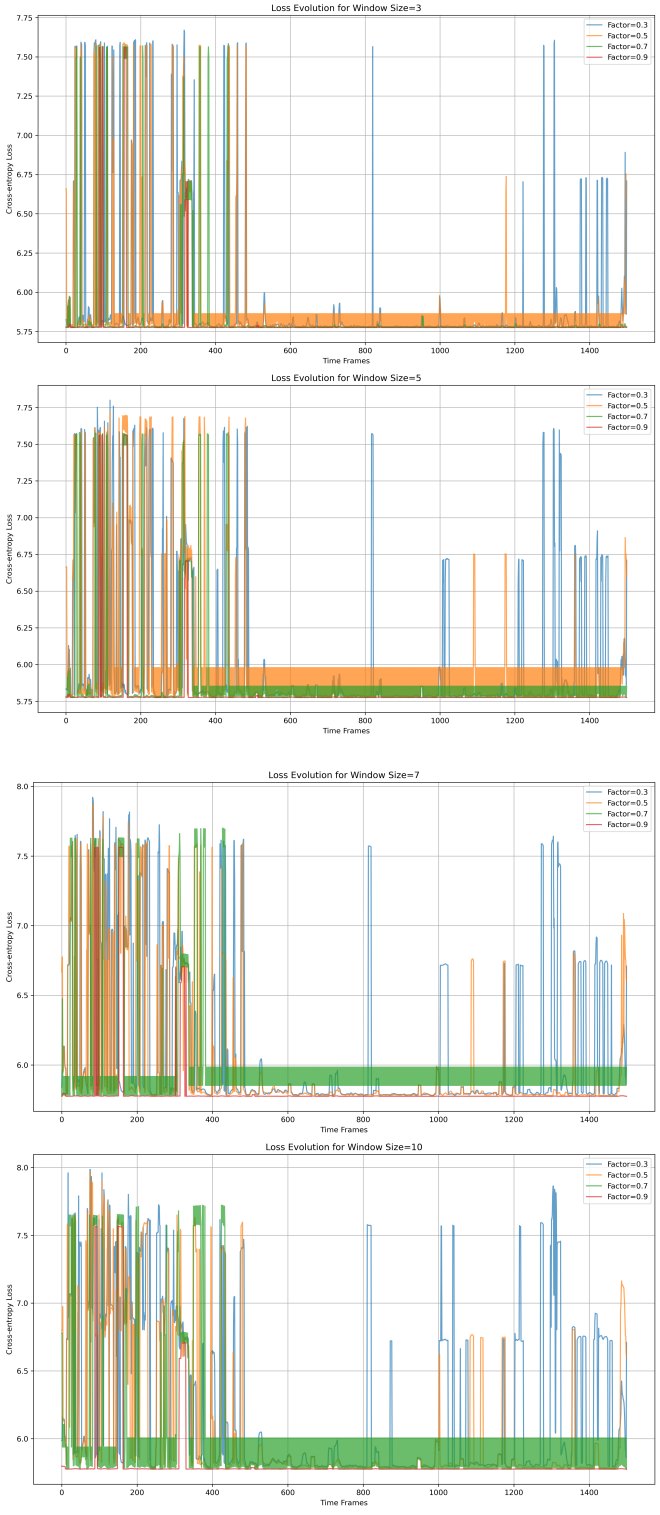

Fig. 8: Cross entropy loss over time frames for different suppression factors for a particular window size.

burden of backpropagation that is typical in many attention-based analysis methods.

Through comprehensive experiments varying suppression factors and window sizes, we uncovered several key insights about Whisper's behavior. We identified which temporal segments had the strongest influence on both individual token generation and complete transcriptions. Our analysis revealed that speech segments exhibited substantially higher influence compared to silent or padded regions. At the token level, we found that the model's predictions were predominantly influenced by immediately preceding temporal frames, indicating a localized temporal dependency in the attention mechanism. Furthermore, our results demonstrated that increasing both the suppression window size and the aggressiveness of suppression factors led to greater disruption in model predictions, with more aggressive suppression (lower factors) causing higher cross-entropy loss. This enables us to provide local explainability for the model's behavior in token generation and elucidate the relationship between specific audio time frames and the resulting transcription.

B. Limitations

While the adaptation of AtMan to Whisper ASR looks promising, there are still limitations to the approach. The observed limitations of the approach are noted below:

1) The adaptation is limited to temporal resolution, making frequency-wise resolution unattainable. This limitation arises because attention manipulation is performed in the encoder layers, which do not allow for frequency-specific adjustments. AtMan's perturbation method, which manipulates the attention, can only perturb over temporal frames.
2) Although we can establish a relationship along the temporal axis by noting that the 2nd Convolutional layer in the encoder reduces the temporal axis by half with a stride of 2 and a kernel size of 3, it is not possible to interpret frequency-wise attributions through attention manipulation due to the Whisper architecture's method of processing audio input. The log mel spectrogram, serving as the "audio input," undergoes two 1D Convolutions that preserve time-wise locality. The embeddings fed into the attention layer maintain a time-wise correspondence to the log-mel spectrogram, but the 80 mel channels are transformed into 384 embedding channels. Each of these 384 features for each time frame is a combination of all 80 mel channels.
3) During token-level analysis, we observed that temporal frames distant from the target token showed unexpected influence on predictions, manifesting as elevated cross-entropy loss. We hypothesize this may be due to token-to-timeframe alignment issues. Specifically, when tokens are split into multiple sub-tokens during processing, calculating loss at each temporal index can result in token position shifts, potentially leading to artificially high loss values in seemingly unrelated temporal regions.
4) The evaluation was conducted on a limited number of samples from the LibriSpeech dataset. A more compre-

hensive evaluation using a larger sample size would provide better insights into the model's behavior. Additionally, while we used Whisper's 'tiny' variant, evaluating other model sizes (base, small, medium, large) would help understand if these findings generalize across model scales.

5) The current analysis focused solely on English language audio. Given Whisper's multilingual capabilities, extending this evaluation to other languages would help validate whether these explainability patterns hold across different linguistic contexts.
6) This generates a local explanation map for each token in the transcription and for whole transcription over the temporal axis. It would be interesting to generate a global explanation map for the model's behavior, potentially by averaging the local explanation maps of all tokens or at sentence level.

C. Future Work

Building upon our current findings, several promising directions for future research emerge:

1) Analyze frequency-specific attributions, possibly by manipulating the convolutional layers or developing hybrid approaches that combine attention and feature-map manipulation.
2) Address the token alignment challenges by developing more sophisticated methods for mapping between temporal frames and output tokens, potentially incorporating forced alignment techniques.
3) Expand the analysis to Whisper's multilingual capabilities, investigating how attention patterns and temporal dependencies vary across different languages and acoustic features.
4) Apply this methodology across different ASR architectures to understand how various model designs influence temporal dependencies and attention patterns.

REFERENCES

[1] B. Deiseroth, M. Deb, S. Weinbach, M. Brack, P. Schramowski, and K. Kersting, "Atman: Understanding transformer predictions through memory efficient attention manipulation," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36, Advances in Neural Information Processing Systems. Curran Associates, Inc., 2023, pp. 63 437–63 460. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/c83bc020a020cdeb966ed10804619664-Paper-Conference.pdf

[2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[3] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[4] A. Akman and B. W. Schuller, "Audio explainable artificial intelligence: A review," *Intelligent Computing*, vol. 3, p. 0074, 2024. [Online]. Available: https://spj.science.org/doi/abs/10.34133/icomputing.0074

[5] T. Sun, H. Chen, G. Hu, L. He, and C. Zhao, "Explainability of speech recognition transformers via gradient-based attention visualization," *IEEE Transactions on Multimedia*, vol. 26, pp. 1395–1406, 2024.

[6] X. Wu, P. Bell, and A. Rajan, "Explanations for automatic speech recognition," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[7] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[8] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Oct. 2019. [Online]. Available: http://dx.doi.org/10.1007/s11263-019-01228-7

[9] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017.

[10] A. Akman and B. W. Schuller, "Atthear: Explaining audio transformers using attention-aware nmf," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 7015–7019.

[11] V. Haunschmid, E. Manilow, and G. Widmer, "audiolime: Listenable explanations using source separation," *arXiv preprint arXiv:2008.00582*, 2020.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[13] S. Sivasankaran, E. Vincent, and D. Fohr, "Explaining deep learning models for speech enhancement," in *INTERSPEECH 2021*, 2021.

[14] A. Frommholz, F. Seipel, S. Lapuschkin, W. Samek, and J. Vielhaben, "Xai-based comparison of audio event classifiers with different input representations," in *Proceedings of the 20th International Conference on Content-Based Multimedia Indexing*, ser. CBMI '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 126–132. [Online]. Available: https://doi.org/10.1145/3617233.3617265

[15] J. Parekh, S. Parekh, P. Mozharovskyi, G. Richard, and F. d'Alché Buc, "Tackling interpretability in audio classification networks with non-negative matrix factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1392–1405, 2024.

[16] Y. Sun, H. Chockler, X. Huang, and D. Kroening, "Explaining image classifiers using statistical fault localization," in *European conference on computer vision*. Springer, 2020, pp. 391–406.

[17] H. Chockler, D. Kroening, and Y. Sun, "Explanations for occluded images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1234–1243.

[18] X. Wu, P. Bell, and A. Rajan, "Can we trust explainable ai methods on asr? an evaluation on phoneme recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 296–10 300.

[19] G. Javadi, K. A. Yuksel, Y. Kim, T. C. Ferreira, and M. Al-Badrashiny, "Word-level asr quality estimation for efficient corpus sampling and post-editing through analyzing attentions of a reference-free metric," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2024, pp. 863–867.

[20] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, and S. Watanabe, "End-to-end speech recognition: A survey," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 325–351, 2024.

[21] H. N. Bicer, P. Götz, C. Tuna, and E. A. P. Habets, "Explainable acoustic scene classification: Making decisions audible," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2022, pp. 1–5.

[22] V. N. Vitale, F. Cutugno, A. Origlia, and G. Coro, "Exploring emergent syllables in end-to-end automatic speech recognizers through model explainability technique," *Neural Computing and Applications*, pp. 1–27, 2024.

[23] Y. K. Singla, J. Shah, C. Chen, and R. R. Shah, "What do audio transformers hear? probing their representations for language delivery & structure," in *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, 2022, pp. 910–925.

[24] R. Kashefi, L. Barekatain, M. Sabokrou, and F. Aghaeipoor, "Explainability of vision transformers: A comprehensive review and new perspectives," *arXiv preprint arXiv:2311.06786*, 2023.

[25] H. Chefer, S. Gur, and L. Wolf, "Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 387–396.

[26] M. V. Ntrougkas, N. Gkalelis, and V. Mezaris, "T-tame: Trainable attention mechanism for explaining convolutional networks and vision transformers," *IEEE Access*, vol. 12, pp. 76 880–76 900, 2024.

[27] Y. Niu, M. Ding, M. Ge, R. Karlsson, Y. Zhang, A. Carballo, and K. Takeda, "R-cut: Enhancing explainability in vision transformers with relationship weighted out and cut," *Sensors*, vol. 24, no. 9, p. 2695, 2024.

[28] W. Bousselham, A. Boggust, S. Chaybouti, H. Strobelt, and H. Kuehne, "Legrad: An explainability method for vision transformers via feature formation sensitivity," *arXiv preprint arXiv:2404.03214*, 2024.
[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
[30] A. Ali, T. Schnake, O. Eberle, G. Montavon, K.-R. Müller, and L. Wolf, "XAI for transformers: Better explanations through conservative propagation," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162, Proceedings of the 39th International Conference on Machine Learning. PMLR, 17–23 Jul 2022, pp. 435–451. [Online]. Available: https://proceedings.mlr.press/v162/ali22a.html
[31] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.
[32] P. Symons, *Digital waveform generation*. Cambridge University Press, 2013.
[33] L. R. Rabiner, R. W. Schafer *et al.*, "Introduction to digital speech processing," *Foundations and Trends® in Signal Processing*, vol. 1, no. 1–2, pp. 1–194, 2007.
[34] "Mel scale," https://en.wikipedia.org/wiki/Mel_scale, accessed: 2023-10-05.
[35] C. Manning, "Lecture 6: Speech recognition and synthesis," https://nlp.stanford.edu/courses/lsa352/lsa352.lec6.6up.pdf, 2023, accessed: 2023-10-05.
[36] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.
[37] A. Mao, M. Mohri, and Y. Zhong, "Cross-entropy loss functions: Theoretical analysis and applications," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 23 803–23 828. [Online]. Available: https://proceedings.mlr.press/v202/mao23b.html
[38] E. Gordon-Rodriguez, G. Loaiza-Ganem, G. Pleiss, and J. P. Cunningham, "Uses and abuses of the cross-entropy loss: Case studies in modern deep learning," in *Proceedings on "I Can't Believe It's Not Better!" at NeurIPS Workshops*, ser. Proceedings of Machine Learning Research, J. Zosa Forde, F. Ruiz, M. F. Pradier, and A. Schein, Eds., vol. 137. PMLR, 12 Dec 2020, pp. 1–10. [Online]. Available: https://proceedings.mlr.press/v137/gordon-rodriguez20a.html
[39] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

ACKNOWLEDGMENT

I extend my sincere gratitude to my supervisors, Prof. Dr. Sebastian Houben and Roman Bartolosch, whose mentorship and expertise were instrumental in shaping this project. I am grateful to Fraunhofer-Institut für Kommunikation, Informationsverarbeitung und Ergonomie (FKIE) for providing the research environment, facilities and support that made this work possible. Finally, I am deeply thankful to my family and friends who have been an unwavering source of support, encouragement and motivation throughout this academic journey.

STATEMENT OF ORIGINALITY

I, the undersigned below, declare that this work has not previously been submitted to this or any other university and that it is, unless otherwise stated, entirely my own work. The report was, in part, written with the help of the AI assistant 'GPT-4o' for the task of copy writing and formatting the report. I am aware that content generated by AI systems is no substitute for careful scientific work, which is why all AI-generated content has been critically reviewed by me, and I take full responsibility for it.

Date Signature

APPENDIX

List of Abbreviations

TABLE I: List of Abbreviations used throughout the report

Abbreviation	Full Form
AI	Artificial Intelligence
ASR	Automatic Speech Recognition
AtMan	Attention Manipulation
CAM	Class Activation Mapping
DNN	Deep Neural Network
LIME	Local Interpretable Model-agnostic Explanations
NLP	Natural Language Processing
NMF	Non-negative Matrix Factorization
POC	Proof of Concept
SHAP	SHapley Additive exPlanations
STFT	Short-Time Fourier Transform
ViT	Vision Transformer
XAI	Explainable Artificial Intelligence

Code

The code for this project is available at https://github.com/arunimaCh29/AtMan-Whisper.

Additional Example Visualizations from the Dataset

Additional samples from the dataset are presented below, illustrating various explainability maps at both sentence and token levels. These examples further support the insights presented in the main report.

Transcription 1: *Well now Ennis I declare you have a head and so has my stick*

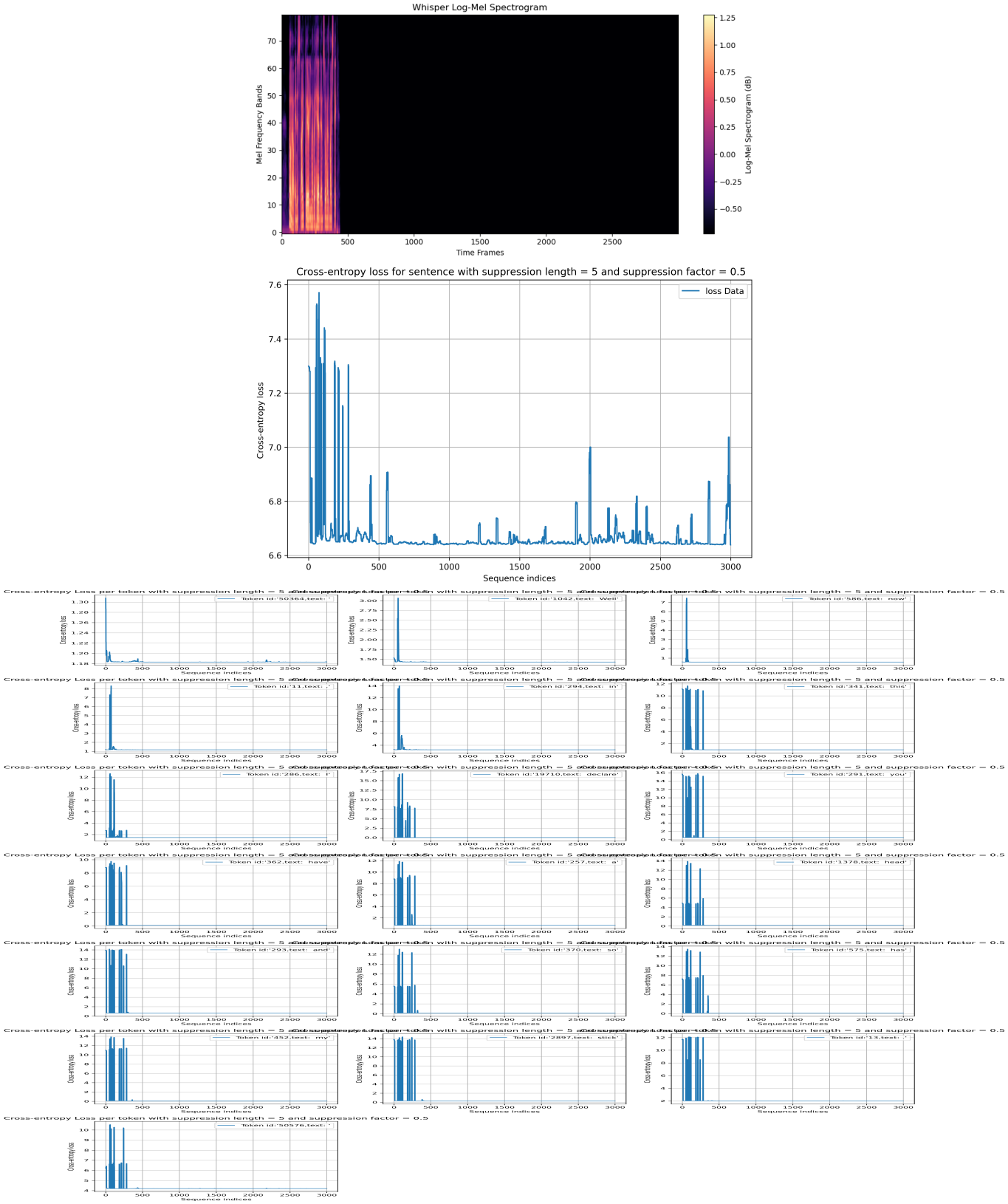

Fig. 9: Visualization set 1: (top) Log Mel spectrogram, (middle) sentence-level influence map, (bottom) token-level influence maps

Transcription 2: *On Saturday mornings when the sodality met in the chapel to recite the little office his place was a cushioned kneeling desk at the right of the altar from which he led his wing of boys through the responses*

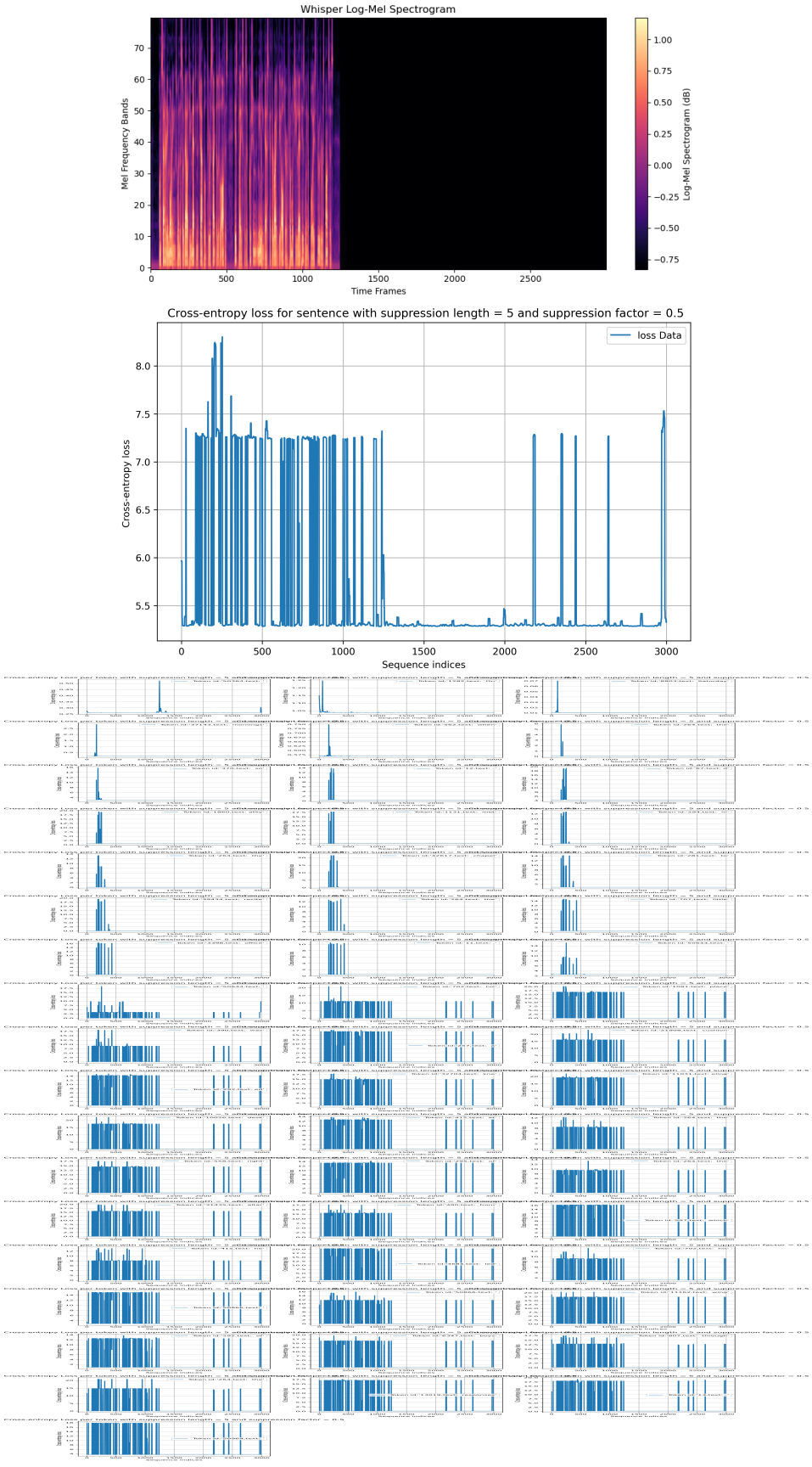

Fig. 10: Visualization set 2 (continued): token-level influence maps

Transcription 3: *Her eyes seemed to regard him with mild pity her holiness a strange light glowing faintly upon her frail flesh did not humiliate the sinner who approached her*

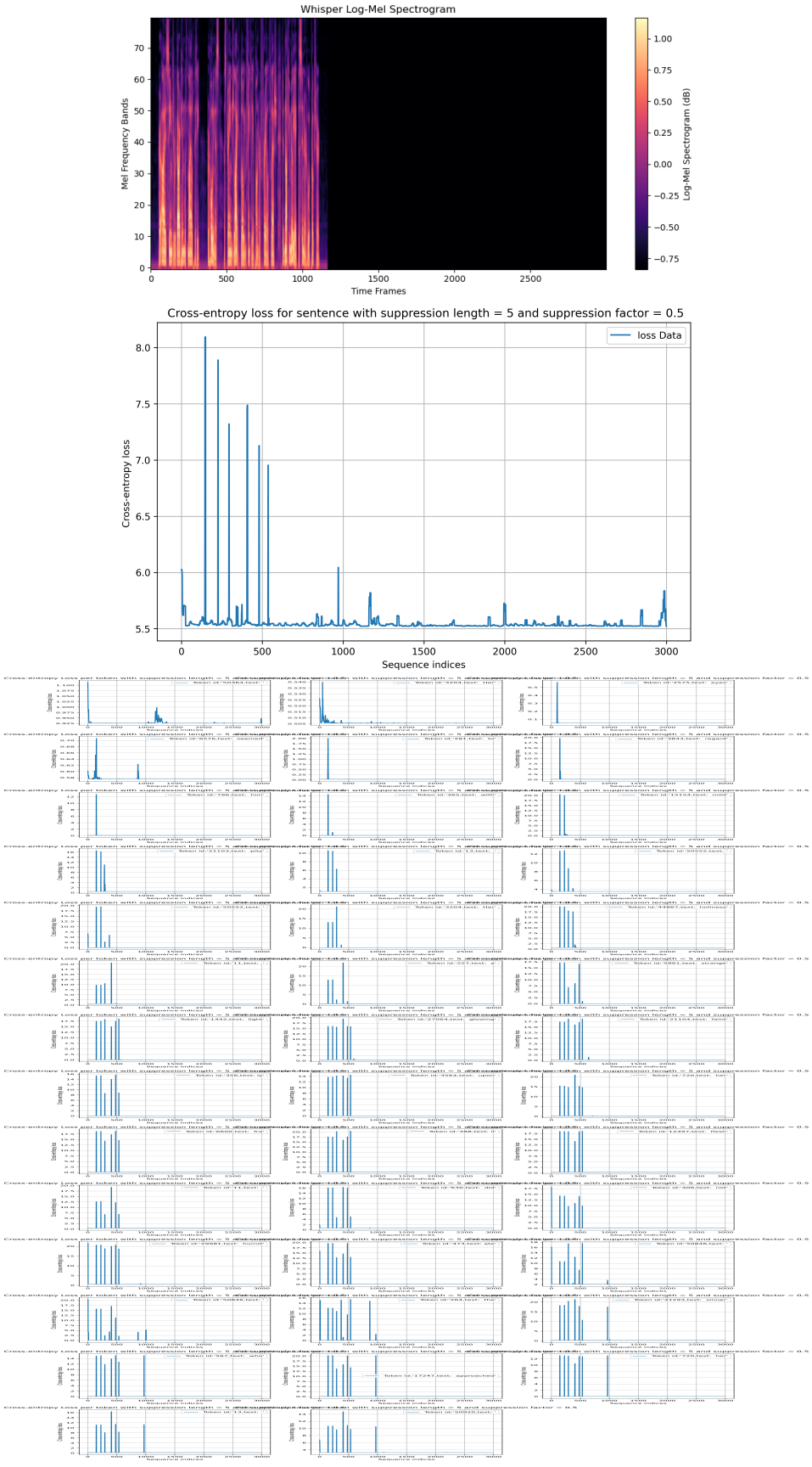

Fig. 11: Visualization set 3 (continued): token-level influence maps

Transcription 4: *If ever he was impelled to cast sin from him and to repent the impulse that moved him was the wish to be her knight*

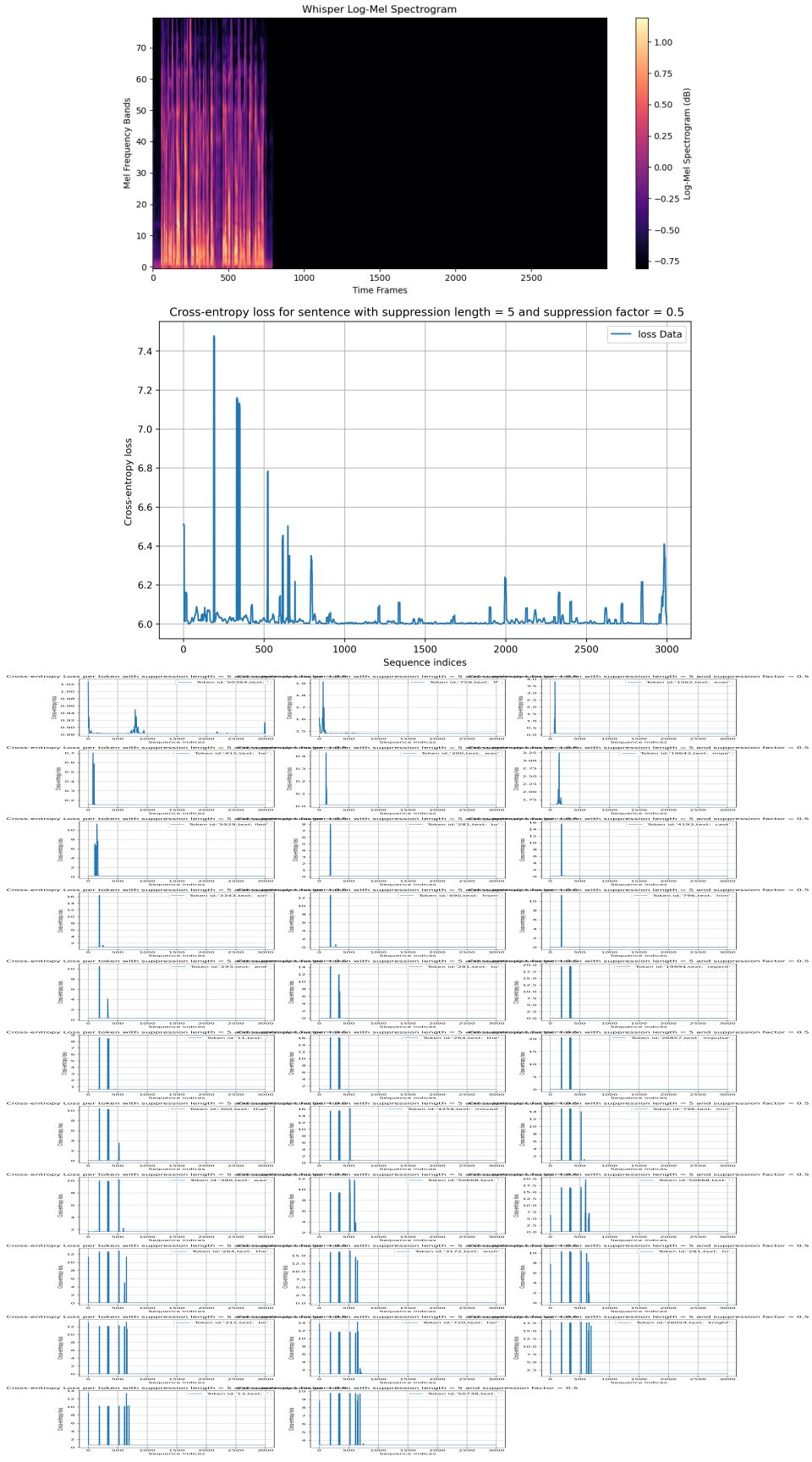

Fig. 12: Visualization set 4: (top) Log Mel spectrogram, (middle) sentence-level influence map

Transcription 5: *He tried to think how it could be*

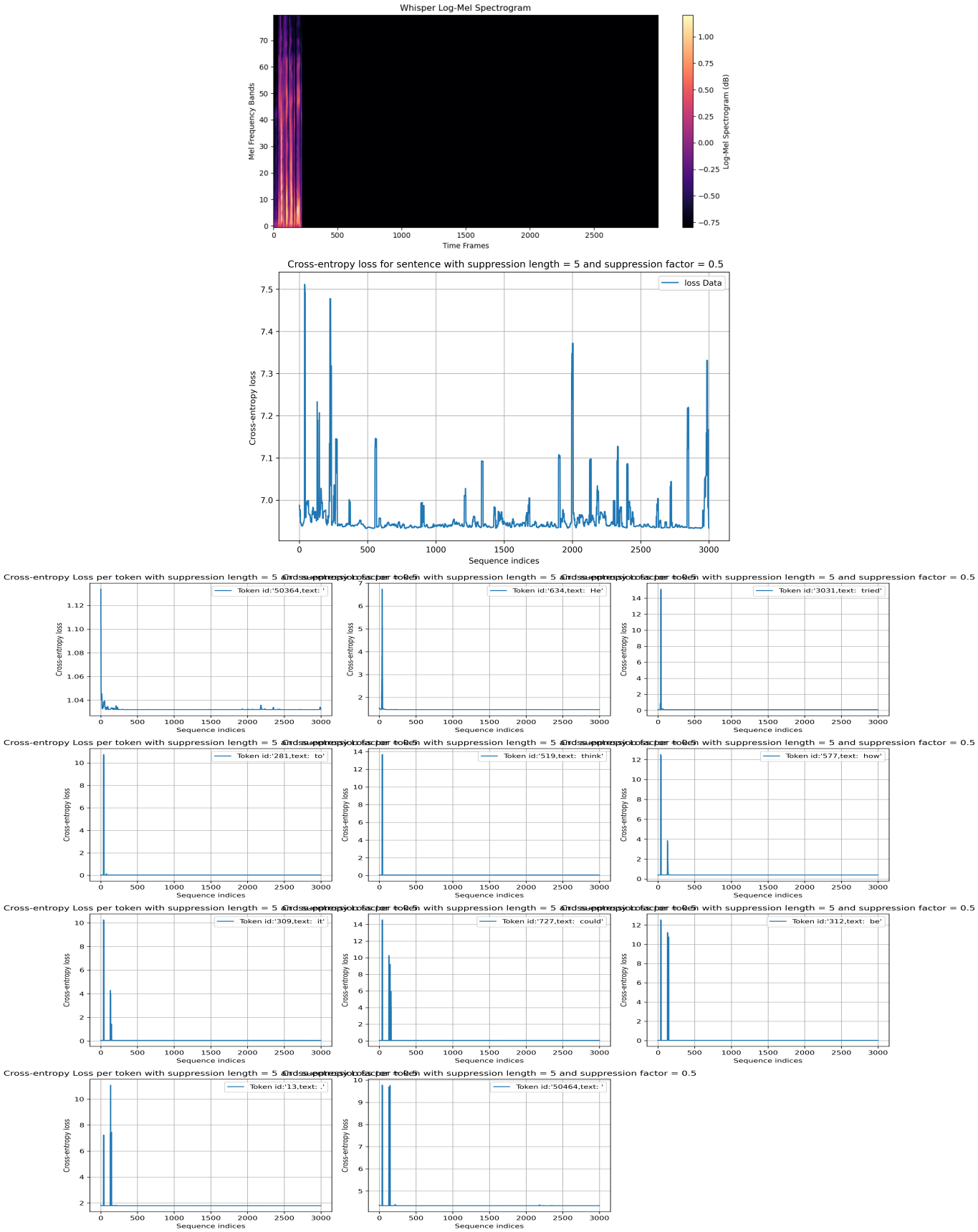

Fig. 13: Visualization set 5: (top) Log Mel spectrogram, (middle) sentence-level influence map, (bottom) token-level influence maps