Soundscape Captioning using Sound Affective Quality Network and Large Language Model

Yuanbo Hou, Qiaoqiao Ren, Andrew Mitchell, Wenwu Wang, Jian Kang, Tony Belpaeme, Dick Botteldooren

***Abstract*—We live in a rich and varied acoustic world, which is experienced by individuals or communities as a *soundscape*. Computational auditory scene analysis, disentangling acoustic scenes by detecting and classifying events, focuses on objective attributes of sounds, such as their category and temporal characteristics, ignoring their effects on people, such as the emotions they evoke within a context. To fill this gap, we propose the affective soundscape captioning (ASSC) task, which enables automated soundscape analysis, thus avoiding labour-intensive subjective ratings and surveys in conventional methods. With soundscape captioning, context-aware descriptions are generated for soundscape by capturing the acoustic scenes (ASs), audio events (AEs) information, and the corresponding human affective qualities (AQs). To this end, we propose an automatic soundscape captioner (SoundSCaper) system composed of an acoustic model, i.e. SoundAQnet, and a large language model (LLM). SoundAQnet simultaneously models multi-scale information about ASs, AEs, and perceived AQs, while the LLM describes the soundscape with captions by parsing the information captured with SoundAQnet. SoundSCaper is assessed by two juries of 32 people. In expert evaluation, the average score of SoundSCaper-generated captions is slightly lower than that of two soundscape experts on the evaluation set D1 and the external mixed dataset D2, but not statistically significant. In layperson evaluation, SoundSCaper outperforms soundscape experts in several metrics on datasets D1 and D2. In addition to human evaluation, compared to other automated audio captioning (AAC) systems with and without LLM, SoundSCaper performs better on the ASSC task in several natural language processing (NLP) based metrics. Overall, SoundSCaper performs well in human subjective evaluation and various objective captioning metrics, and the generated captions are comparable to those annotated by soundscape experts. The model, source code, LLM scripts, human assessment data, instructions, and evaluation statistics are all publicly available.**

***Index Terms*—Soundscape, acoustic scene, audio event, affective quality, large language model, soundscape caption**

I. Introduction

Soundscape plays a vital role in shaping our daily experience, affecting various aspects including mood, behaviours, and overall well-being [1]. Understanding how people perceive soundscape has become important in urban design, environmental psychology, and interactive technology [2]. The international standard organization (ISO) 12913-1:2014 [3] defines soundscape as: "*the acoustic environment as perceived or experienced and/or understood by a person or people, in context*", which emphasizes the interaction between a person or people and the acoustic environment. Thus, a soundscape is not only about the audio events (AEs) present in the acoustic scene (AS), but also the emotions and mood evoked by the acoustic environment and events therein.

To describe the affect of soundscape, ISO/TS 12913-3:2019 [4] recommends using the soundscape circumplex model (SCM) [5], a framework inspired by the affect theory of emotions [6]. In Fig. 1, the SCM is scored on eight 5-point Likert scales to describe the perceptual attributes of soundscapes. Some prior studies [7] [8] explore the relationships between AEs and annoyance, which is one of the 8 attributes of perceived affective quality (PAQ) in SCM. These perceptions are shaped by sound characteristics, contextual cues, and prior experience or common knowledge specific to culture. In addition, the affect of soundscape is highly related to the perception of AS as a whole. Therefore, to describe a soundscape, it is crucial to exploit both the physical information about the acoustic environment that can be estimated using techniques such as acoustic scene classification (ASC) and audio event classification (AEC), and the perceptual information, such as human-perceived affective quality (AQ) of the soundscape.

To enable machines to understand acoustic environments, computational analysis of audio scenes and events [9] has been studied extensively, e.g., by the detection and classification of acoustic scenes and events (DCASE) community [10]. This has led to significant advancements in the recognition of ASs and AEs, resulting in various methods, from machine learning methods to deep learning methods, such as AE detection methods [10] [11] [12] based on frame-level strong labels or clip-level weak labels, and ASC methods [13] [14] [15]. More recently, natural language has been used to describe audio content, including ASs and AEs, leading to an emerging area called automated audio captioning (AAC) [16] [17].

However, the DCASE-related works focus mainly on the

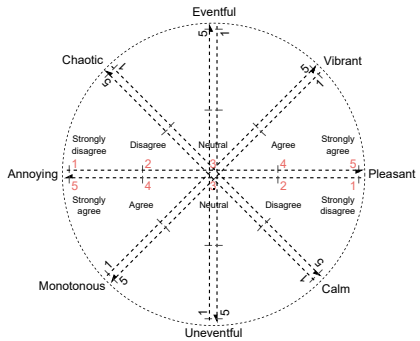


Fig. 1. The 8-dimensional (8D) affective qualities (AQs) in the soundscape circumplex model (SCM) [5] recommended by ISO/TS 12913-3:2019 [4].

This research received funding from Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

Yuanbo Hou is with the Machine Learning Group, University of Oxford, UK. Corresponding email: Yuanbo.Hou@eng.ox.ac.uk

Qiaoqiao Ren and Tony Belpaeme are with the AIRO-IDLab, Ghent University-Imec, Belgium.

Wenwu Wang is with the CVSSP, University of Surrey, Guildford, UK.

Andrew Mitchell and Jian Kang are with University College London, UK.

Dick Botteldooren is with the WAVES Group, Ghent University, Belgium.

objective attributes of sounds, while ignoring the effects these sounds have on people, such as the different emotions they evoke. This results in the relations between various AEs and 8-dimensional (8D) affective qualities (AQs) in PAQ remaining unexplored, let alone the relations between AS, AEs, and affective responses to 8D AQs. For example, the AAC systems often describe AEs or AS-related information, such as "*birds are of chirping the chirping and various chirping*" in the AAC task in DCASE 2020 [17], but they have not explored the listener's response to the audio along the affective dimension, i.e., whether hearing birds chirping brings pleasure or annoyance to the listener. Despite the substantial progress in AAC for the acoustic environment, little attention has been paid to the affective information carried by AEs in a soundscape or the soundscape as a whole. More specifically, there is a significant research gap in captioning that links the acoustic environments (ASs and AEs) and the human-perceived AQs of soundscapes.

To fill this research gap, we propose a new task, called *affective soundscape captioning* (ASSC), where a soundscape is described using context-aware texts, detailing AS, AE, and emotion-related AQ. This enables the exploration of affective information from anthropocentric soundscapes, as defined in ISO 12913-1:2014 [3]. Inspired by recent advancements in large language models (LLMs) [18], we propose an LLM-based soundscape caption system (SoundSCaper) for the ASSC task by integrating coarse-grained ASs, fine-grained AEs, and human-perceived AQs. SoundSCaper integrates the rich prior knowledge in LLMs like generative pre-trained transformer (GPT) models [19], enabling context-sensitive and affectively meaningful descriptions of soundscapes.

To provide precise acoustic inputs to the LLM, we propose a sound affective quality network (SoundAQnet) to simultaneously model AS and AE in acoustic environments, and their corresponding AQ. SoundAQnet combines the Mel-spectrogram features [12] commonly used in ASC and AEC tasks with loudness (Zwicker loudness, defined in ISO 532-1 [20]) to capture AS, AE, and AQ in the soundscape. Loudness, which measures the subjective impression of human perception of sound, has important implications for AQ modelling, as the perceived loudness of a sound can change its contribution to affect, from gentle background to extremely disturbing sounds. Integrating the output from SoundAQnet with GPT allows SoundSCaper to generate natural language descriptions, thereby linking objective acoustic indicators (i.e., AS and AE) with human-centred perceptual indicators (i.e., PAQ). To our knowledge, we are the first to build models that simultaneously characterize AS, AE, and AQ in a soundscape and to describe the soundscape with affective captions using an LLM based on the three-view information (i.e., AS, AE, and AQ).

This paper strives to advance machine listening by linking it with affective computing and contextual interpretation, thus going beyond conventional recognition and classification of sounds. Our work offers the potential to enable machines to have a comprehensive and emotionally attuned perception of auditory scenes and events. The contributions of this paper are as follows: 1) We propose the ASSC task, where a soundscape is described in free texts from the perspectives of AS, AE, and AQ, thus bridging the gap between audio captions and the human-perceived AQs of sounds. 2) We propose SoundAQnet to simultaneously model the coarse-grained AS and fine-grained AE, as well as human-perceived AQ. 3) We utilize the rich knowledge embedded in LLM about expected sounds in various scenes to develop the automatic SoundSCaper system, which translates SoundAQnet's predictions into human-understandable natural language captions; to this end, careful soundscape-focused prompt engineering is introduced. This transforms soundscape descriptions from limited numerical values into comprehensible free text rich in acoustic context and human AQ perception. 4) To measure the quality of the soundscape captions, we introduce the Transparent Human Benchmark for Soundscapes (THumBS) as a metric and evaluate the performance of SoundSCaper on the test set and the mixed external dataset.

Next, Section II discusses related work. Section III introduces the proposed ASSC task. Section IV proposes the SoundSCaper based on SoundAQnet and LLM. Section V presents SoundAQnet experiments. Section VI presents human evaluation experiments on SoundSCaper by expert and layperson groups, compares the results of SoundSCaper and AAC systems on the ASSC task, and discusses their characteristics. Section VII concludes. We have released the code and models, human assessment data, and human evaluation statistics to the ***homepage*** (https://github.com/Yuanbo2020/SoundSCaper).

II. Related Work

This section reviews related work on soundscape captioning.

A. Audio Captioning

Audio captioning [16] [17] resembles soundscape captioning when it comes to disentangling the auditory scene into separate sounds. Various automated audio captioning (AAC) systems [21]–[23] aim to describe AEs and physical properties of acoustic environments using text. ConvNeXt-Trans [22], the baseline of DCASE 2024 challenge [17], excels at the AAC task. The audio encoder of ConvNeXt-Trans is ConvNeXt, and its decoder consists of a Transformer decoder with a structure similar to GPT. P-LocalAFT [23] allows local information to be captured while retaining global information, which captures AEs of different durations for precise captions. These AAC systems focus on AEs and ASs, as a result, they can describe the objective information of acoustic environments. However, they fail to capture emotions and moods perceived by humans, as the datasets used for model training lack AQ labels.

In addition to AAC systems without LLM, audio-LLM-based GAMA [24] and Qwen-Audio [25] perform well on multiple audio captioning datasets. GAMA uses an acoustic model to extract the information and then feeds it to an LLM to generate captions. GAMA's framework is similar to SoundSCaper. Although audio-LLMs [24] [25] have some audio perception capability, such perception focuses on the objective content of audio and serves audio understanding and reasoning; i.e., it is not the affective perception in soundscapes, let alone the perception of 8D AQs in ISO/TS 12913-3:2019.

B. Affective Computing

In soundscape affective computing, arousal-valence dimensional models (AVDM) [2] are often used to capture human

emotional responses. However, AVDM has limited ability to distinguish different affective perceptions, making it difficult to model complex emotion states [26]. Based on the pleasantness-eventfulness framework, the performance of multiple acoustic features on urban soundscapes is analyzed, and the results [27] show that Mel-based features predict the pleasantness and eventfulness of soundscapes well. And Gammatone cepstral coefficients [28] have been shown to be feasible in assessing the valence and arousal of sounds. However, the study [28] is based on synthetic rather than real datasets. This paper models human affective responses based on a multi-user annotated real soundscape dataset according to 8D AQs in ISO standards.

C. Soundscape Analysis

Soundscape analysis can be roughly categorized into acoustic environment-oriented [9]–[12] and affective perception-oriented [1] [5] [29]–[31]. The former focuses on the *objective* acoustic environment understanding, e.g., ASC and AEC. The latter focuses on affective perceptions of soundscapes to characterize emotional states. The international affective digital sound (IADS) [29] dataset explores discrete emotional categories elicited by 167 individual stimuli, and the international soundscape database (ISD) [30] explores the relationship between annoyance and 24 categories of AEs in daily life. In addition, ARAUS [32] studies the unique subjective perceptual responses of 605 participants to 25440 augmented soundscapes presented as audiovisual stimuli.

To the best of our knowledge, there are no studies on automatic descriptions of soundscapes, especially AASC. The successful application of the proposed AASC will promote automated soundscape analysis, improve urban soundscape planning, and improve environmental awareness of visually and hearing-impaired people [33] [34].

III. Affective Soundscape Captioning

This section presents problem formulation and task definition for ASSC, and the SoundScaper system for AASC.

A. Problem Formulation and Task Definition

Affective soundscape captioning (ASSC) describes the holistic experience and understanding of an acoustic environment. Identifying and recognizing AEs and assigning meaning to them is one of the cornerstones of this personal and individual experience. Because of this, soundscape evaluation questionnaires (e.g., ISO/TS 12913-2 [35]) inquire about the classes of sounds people hear. Thus, ASSC should include a description of the relevant categories of audible sounds.

The holistic evaluation is also affected by psychoacoustics-related loudness [20] and the context of sounds. Therefore, from the perspective of holistic experience and understanding, ASSC should consider the categories of sounds, the context in which they occur, and the AQ they may evoke in humans. Fig. 2 contrasts ASSC to related tasks. The AEC task aims to tag the types of AEs of audio clips with semantic labels; the ASC task identifies the environment category where the sound is recorded, i.e., its context. The AAC task converts audio content, mainly AEs, into text. ASSC starts from the perspective of soundscape with human perception rather than just AEs, adds AS information, and focuses on emotional impact. AASC adds affective attributes to the textual description and suggests the human affect evoked by the acoustic environment.

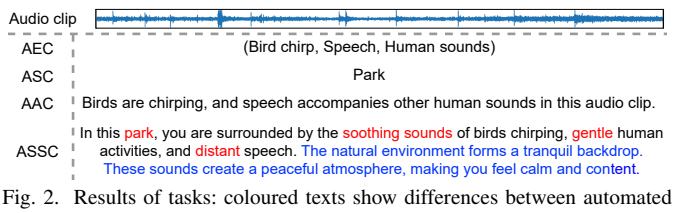


Fig. 2. Results of tasks: coloured texts show differences between automated audio captioning (AAC) and affective soundscape captioning (ASSC) tasks. Blue texts indicate *human-perceived AQ*-related descriptions unique to ASSC.

B. Overview of the Proposed SoundScaper System

The proposed automatic soundscape captioner, SoundSCaper, as shown in Fig. 3, consists of two parts: the acoustic model ($am(\cdot)$) SoundAQnet, and the language model ($lm(\cdot)$). SoundSCaper aims to generate a language description $\boldsymbol{D}$ to describe the soundscape based on the input audio clip $\boldsymbol{A}$. To this end, first, we extract the AS and AE information, as well as the affective response values of eight AQs in Fig. 1, i.e. *pleasant*, *eventful*, *chaotic*, *vibrant*, *uneventful*, *calm*, *annoying*, and *monotonous* [32] from the audio clip $\boldsymbol{A}$, by building an acoustic model ($am(\cdot)$), i.e., $\{AS, AE, AQ\} = am(\boldsymbol{A})$. Then, we form a textual description of the soundscape by a language model like GPT [18] ($lm(\cdot)$), i.e., $\boldsymbol{D} = lm(AS, AE, AQ)$.

IV. The Proposed SoundSCaper System

This section introduces SoundSCaper in detail from two aspects: the acoustic model and the language model.

A. Acoustic Model: the Proposed SoundAQnet

The SoundAQnet aims to simultaneously model ASs and AEs in acoustic environments, as well as the corresponding affective responses to the soundscape, i.e., PAQ 8 attributes shown in Fig. 1. In recognition of AS and AE, the Mel spectrogram is a typical acoustic feature [12] [15] with excellent performance. In soundscape studies [1] [2] [31], perceived loudness, approximated as calculated as Zwicker-loudness in ISO 532-1:2017 [20], is of primary importance to estimate AQ even if AEs are known. Thus, both Mel and loudness are used to capture the AS, AE, and AQ in soundscape audio clips. The novelty of SoundAQnet lies in the design of the network structure and the loss function to enable the joint modelling of ASs, AEs, and AQs, as described below.

1) ***Network:*** To process these long input features with few parameters, SoundAQnet uses dilated convolution [36] in its Mel-based and Loudness-based branches to obtain a larger receptive field size (RFS) with limited computing resources.

1.1) ***Mel-based Branch:*** Audio events are often in different spectral-temporal scales, ranging from short and transient to long-lasting events. Therefore, the Mel-based branch employs

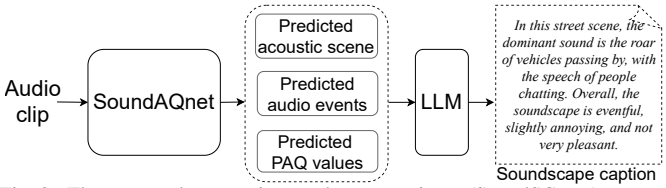


Fig. 3. The proposed automatic soundscape captioner (SoundSCaper) system.

four sub-branches to obtain multi-scale representation, by applying convolutional kernels of different sizes, i.e., [(3, 3), (5, 5), (7, 7), (9, 9)], to the input features on the (time, frequency) axis, respectively. Each sub-branch comprises three convolution blocks with 16, 32, and 64 filters. The gridding artifacts [37] of dilated convolutions can lead to compromised information continuity and loss of local information. Thus, a hybrid dilated convolution [36] scheme is used, where the dilation rates in the three convolutional blocks are in order [(1,1), (2,1), (3,1)], allowing the branches to extract context from a broader and more coherent receptive field along the time axis. The dilation rate only varies along the time axis, as the frequency dimension is often relatively small.

In the Mel branch, each 2D convolution (Conv2D) block refers to the design of VGG [38] and consists of two convolution layers. Taking the largest kernel (9, 9) as an example, there are three Conv2D blocks, i.e., six 2D convolution layers, according to the convolution RFS calculation formula,

$$F_i = (F_{i-1} - 1) \times stride + k \quad (1)$$

where F_i denotes the i-th convolution layer's RFS relative to the input feature map, $F_0 = 1$, k is the convolution kernel size, and $stride$ defaults to 1. If there is no pooling operation, according to Eq. (1), in the first Conv2D block, the RFS of the first layer on the time axis is $F_1 = 9$, and that of the second layer is $F_2 = 17$. For the dilated convolution, the RFS is

$$F_i = (F_{i-1} - 1) \times stride + k + (k-1)(r-1) \quad (2)$$

where r is the dilation rate. For the second Conv2D block with dilation rate (2, 1), on the time axis, $F_3 = 33$, and $F_4 = 49$. For the third Conv2D block with dilation rate (3, 1), on the time axis, $F_5 = 73$, and $F_6 = 98$. With these Conv2D blocks without pooling, SoundAQnet requires the input features to be at least 98 frames long. With the frame hop of 10*ms*, the corresponding input clip length is at least 980*ms*. It is challenging to identify AS or AE from 1-second audio clips, even for humans, let alone the 8D AQs. Furthermore, if pooling is not used in Conv2D, it will increase the parameters and the computation load. After comprehensive trade-offs, we add pooling operations to these multiscale Conv2D blocks, resulting in a minimum input audio length of 2.80*s*.

After the last Conv2D block of each sub-branch, global pooling unifies the length of multiscale representations. These dimension-unified representations are fed into separate embedding layers to output 64-dimensional embeddings for fusion.

1.2) ***Loudness-based Branch****:* The loudness of AEs affects how humans perceive them, thereby affecting human-perceived AQ. Many models for noise annoyance within one class of sounds (e.g., road traffic) even rely exclusively on loudness. The Mel spectrogram can estimate the loudness changes to some extent. However, since perceived loudness has been studied in psychoacoustics for decades, it is more effective to directly use loudness as a psychoacoustic feature in the model instead of estimating it implicitly with the model. Loudness is introduced as a 1D feature with the unit of **sone**. In SoundAQnet, loudness features are extracted after calibration with a reference signal, i.e. a sine wave of 1kHz at 60dB, from the ISO 532-1 standard [20]. Given N frames, the size of loudness features is (N, 1). Since the generation, development, and fading of emotions is a dynamic process with different time scales of importance, we also use multiscale convolution blocks in the loudness branch to extract the PAQs defined in ISO/TS 12913-3:2019 [4] shown in Fig. 1. More specifically, the multiscale convolution kernels used in the sub-branch are of dimension [(3, 1), (5, 1), (7, 1), (9, 1)]. The remaining part of the loudness branch is the same as that in the Mel branch.

1.3) ***Graph-based Multiscale Embedding Fusion****:* To fuse representations from Mel- and Loudness-based branches, we consider the representation embeddings as node features and build a fully connected soundscape-dependent multiscale sound-AQ representation graph. Here, our hypothesis is that since the model is trained with co-supervised labels of AS, AE, and AQ in the soundscape, the sound-AQ representation graph will automatically couple the acoustic environment and AQ while updating node features and learning relationships between nodes with different time granularities. That is, by updating the features of edges connecting nodes, the message about the differences between different timescale nodes is passed to each other through edges in the graph, thereby further aggregating and fusing information from different scales. Thus, it is crucial to learn edge features in the sound-AQ representation graph during updating. Then, we use the gated graph convolutional network (GatedGCN) [39] in the graph-based multiscale embedding fusion layer, where the

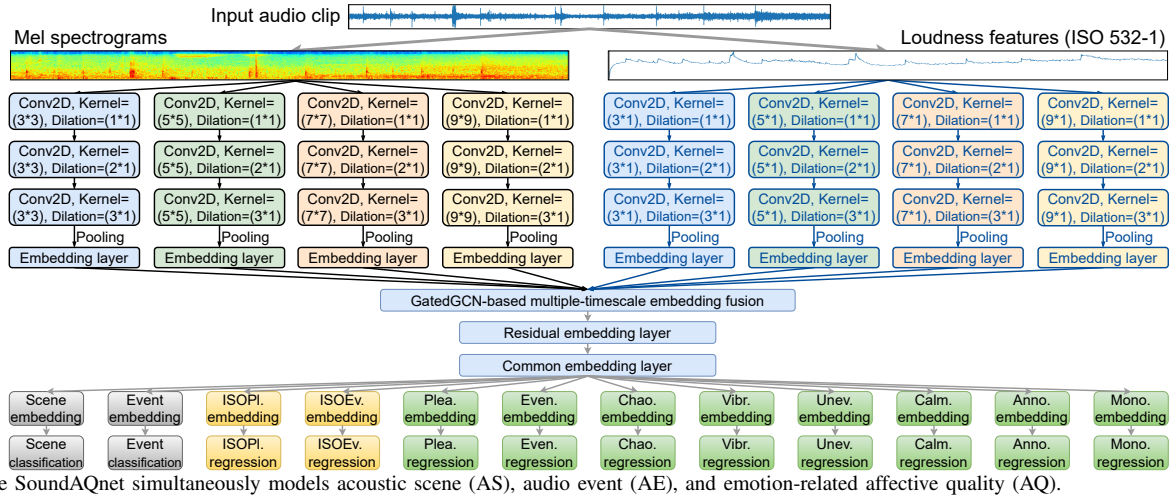


Fig. 4. The SoundAQnet simultaneously models acoustic scene (AS), audio event (AE), and emotion-related affective quality (AQ).

node and edge features are updated simultaneously. Once the soundscape-dependent sound-AQ representation graphs containing n nodes and $n \times n$ edges are obtained, we use one layer of GatedGCN to model these graphs. The number of nodes is $n = 8$; the size of each node embedding is 64.

1.4) ***Co-embedding and Separate Embedding Layers****:* To absorb the information updated by the sound-AQ graph while considering its original acoustic representations, we residually connect [40] the node embeddings output by the graph with the input to the graph. We concatenate all embeddings and input them into the common embedding layer to learn the acoustic context- and AQ-related embeddings. This allows the classification and regression tasks to use all the information within the common embedding captured by the model. Next, separate embedding layers are used for the ASC and AEC tasks, and the human-perceived AQ regression tasks, to learn representations for each target individually.

2) ***Loss Functions****:* The SoundAQnet involves 2 classification objectives (AS, AE) and 10 regression objectives (*ISOP*, *ISOE*, 8D AQs). ISO Pleasantness (*ISOP*) and ISO Eventfulness (*ISOE*) in Fig. 1, can be calculated as follows:

$$ISOP = k^{-1}(\sqrt{2}r_{pl} - \sqrt{2}r_{an} + r_{ca} - r_{ch} + r_{vi} - r_{mo}) \quad (3)$$

$$ISOE = k^{-1}(\sqrt{2}r_{ev} - \sqrt{2}r_{ue} - r_{ca} + r_{ch} + r_{vi} - r_{mo}) \quad (4)$$

where $r_{\{pl,ev,ch,vi,ue,ca,an,mo\}} \in \{1, 2, 3, 4, 5\}$ are human response values to 8D AQs: *pleasant*, *eventful*, *chaotic*, *vibrant*, *uneventful*, *calm*, *annoying*, and *monotonous*, respectively, and $k = 8 + \sqrt{32}$. *ISOP* and *ISOE* are related to AQs, so the model's prediction for *ISOP* can imply the overall performance of human-perceived AQ predictions.

2.1) ***AS****:* For the ASC tasks, cross entropy (CE) [15] is used as the loss function that measures the difference between the prediction p_s and its label y_s, i.e. $\mathcal{L}_1 = CE(p_s, y_s)$.

2.2) ***AE****:* For the AEC tasks, binary cross entropy (BCE) is used as the loss function that measures the difference between the prediction p_e and its label y_e, i.e. $\mathcal{L}_2 = BCE(p_e, y_e)$.

2.3) ***AQ****:* For the AQ regression tasks, mean squared error (MSE) is used as the loss function. Specifically, $\mathcal{L}_3 = MSE(p_{isop}, ISOP)$ and $\mathcal{L}_4 = MSE(p_{isoe}, ISOE)$, where p_{isop} and p_{isoe} are predictions of *ISOP* and *ISOE*, respectively. Also, $\mathcal{L}_{\text{n}} = MSE(p_{aq}, y_{aq}), n \in [5, 12]$, where p_{aq} and y_{aq} are predictions and labels of each type of AQ in 8D AQs.

2.4) ***Total Loss****:* Optimizing the 12 objectives with 12 losses is challenging. Typical Pareto optimization [41] is unsuitable for SoundAQnet. Because the quantification of AQ has a certain degree of ambiguity, assuming that $3 \pm 0.25 \approx 3$ for r_{pl}, its prediction ± 0.1 has little impact on the final AQ output. Hence, compared to emotion-related AQs, SoundAQnet needs to perform better in ASC and AEC with explicit classification goals, i.e., SoundAQnet does not aim to achieve the Pareto optimality of all 12 objectives. Human perception times for various scenes, events, and emotions may vary. This implies that different learning rates might be better suited for optimizing the 12 classification and regression losses. Hence, GradNorm-like optimizations [42], which aim to learn multiple tasks at a similar rate from a gradient view, do not suit SoundAQnet. After considering the computational effort and training speed, we choose uncertainty-based weighting [43] to fuse the 12 losses.

$$\mathcal{L} = \sum\nolimits_{i=1}^{2} \left(\frac{1}{{\sigma_i}^2}\mathcal{L}_i + \log\sigma_i\right) + \sum\nolimits_{j=3}^{12} \left(\frac{1}{2{\sigma_j}^2}\mathcal{L}_j + \log\sigma_j\right) \quad (5)$$

where the learnable noise parameter σ denotes the uncertainty associated with the corresponding task [43], and the logarithmic function based penalty term prevents σ from becoming excessively high. The higher the uncertainty σ, the lower the contribution of the loss to the total loss.

B. Language Model: Customised LLM for Caption Generation

This section explains how the AS, AE, and AQ information, obtained by SoundAQnet, can be turned into captions to describe the soundscape in human-understandable language by customizing an LLM with prompt engineering techniques.

1) ***Pipeline for Caption Generation****:* As shown in Fig. 5, we design a pipeline to convert the numerical information of AS and AE and the emotion-relevant AQ into a textual description of the soundscape, with the help of the prior knowledge learned by an LLM. There are various methods for integrating acoustic information into LLM, including early, mid, and late fusion [44], corresponding to information fusion at the feature, intermediate representation, and decision level, respectively. In this work, we choose GPT as the LLM to fuse the decision results of scenes, events, and affective qualities parsed by SoundAQnet, i.e. $\boldsymbol{D} = lm(AS, AE, AQ)$ as discussed in Section III, where lm represents the GPT model, such as DaVinci, GPT-3.5-Turbo, and GPT-4, according to OpenAI [18] services. We choose GPT-3.5-Turbo, which offers a tradeoff among generation accuracy, response speed, number of tokens, and cost. An advantage of this pipeline is that SoundSCaper's pipeline can directly use general LLMs such as GPT and potentially benefit from their future updates.

In the SoundSCaper system, LLM serves two purposes: (1) to present knowledge extracted from audio recordings using SoundAQnet in fluent language; (2) to introduce common knowledge on expected AEs and AQs in a specific context, that is, a specific AS. Currently, there is no large-scale dataset in the soundscape domain that pairs audio with affective descriptive captions encompassing AE, AS, and human-perceived AQ. As a result, existing resources are inadequate for training audio-to-text soundscape-language models or for fine-tuning LLMs. In addition, fine-tuning LLMs on expert-annotated affective captions may introduce biases related to individual personalities and cultural perspectives. Moreover, fine-tuning LLM may increase the chance of hallucination [45]. For these reasons, SoundSCaper uses a general pre-trained GPT instead of fine-tuning a model. This choice has several advantages: it

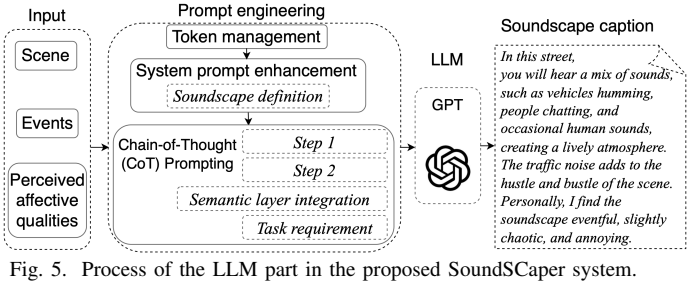


Fig. 5. Process of the LLM part in the proposed SoundSCaper system.

avoids the dilemma of no paired data of sound and affective description to support the training or fine-tuning of soundscape LLM, and also avoids time-consuming and laborious training; directly using generic LLMs allows the acoustic and language models to be updated separately; the system can be extended with non-acoustic contextual information, if available; and captions can be easily provided in other languages.

*2) **Customized LLM**:* As a novel design for soundscape caption generation, we integrate prompt engineering to enhance the output's contextual accuracy, affective depth, and narrative clarity to decrease hallucinations, as shown in Fig. 5. The employed prompt engineering strategies include structured output generation, system prompt enhancement, semantic layer integration, and chain-of-thought (CoT) prompting, which are more effective than fine-tuning LLM in improving model performance and reducing hallucinations [46].

*2.1) **Token Management**:* To optimize the amount of input and output of the LLM, tokens are managed. For an audio clip, the scene, AS, is unique, but multiple AEs may be detected simultaneously. Hence, to process AE probabilities predicted by SoundAQnet, we first use an empirical probability threshold of 0.3 to obtain the text labels of AEs present in the audio clip. Then, for AQs predicted by SoundAQnet, we prioritize input tokens with strong responses, that is, high predicted values. This is because human attention is often attracted by the dominant AE while being influenced by the dominant AQ. The ASSC task aims to describe the most relevant information in soundscapes. Additionally, we instruct the LLM to limit the generated descriptions to 200 tokens. These strategies manage the consumption of input and output tokens.

*2.2) **System Prompt Enhancement**:* We use the soundscape definition from ISO 12913-1:2014 [3] to provide LLM with a conceptual framework that incorporates perception (psychology) and understanding (cognition) in the description, as well as context, people, and society. In addition, the prompt asks the LLM to define its role as a soundscape expert to fit the soundscape research field.

*2.3) **Semantic Layer Integration**:* Rule-based semantic constraints are used to reduce hallucinations [46] in ASSC. The LLM is instructed to use only the provided AE labels and avoid interpretive or embellished descriptions. For example, *vehicle* should not be expressed as 'the hum of engines and honking of horns'; LLM should directly list the AEs. These constraints help ground the output, improving accuracy and consistency without requiring LLM fine-tuning.

*2.4) **Chain-of-thought Prompting**:* This part guides the LLM through a logical analysis sequence for structured output generation, from AS and AE classification to AQ regression. The structured approach aids in systematically tackling complex auditory and affective analyses. The task is decomposed into focused subtasks to help LLM understand the relationship between input acoustic environment information and AQs based on its large-scale prior knowledge to ensure comprehensive and accurate caption generation. Prompts are as follows:

...As an expert in soundscape analysis, your task is ...

Step 1: According to the events and their corresponding probability ... in this scene, identify ... and describe the auditory scenario...

Step 2: Describe your feelings based on the ratings on this soundscape.....

Full prompts and scripts can be found in the ***homepage***.

V. Acoustic model (SoundAQnet) experiment

A. Dataset

Commonly used large-scale audio datasets like AudioSet [47] and FSD50K [48] do not contain corresponding "*subjective*" labels regarding the PAQ of recording environments [32], which prevents them from being used to train SoundAQnet. To the best of our knowledge, the recently published ARAUS dataset [32] is the largest soundscape dataset with the most complete human affective responses to AQs. Therefore, the ARAUS dataset is used to train the proposed SoundAQnet.

ARAUS contains 25440 30-second binaural audio clips, totaling 212 hours. With the efforts of 605 participants, each audio clip has 8D AQ values annotated according to ISO/TS 12913-2 [35]. ARAUS is augmented on the Urban Soundscapes of the World (USotW) [49] dataset. Each augmented soundscape is made by digitally adding maskers (*birds*, *water*, *wind*, *traffic*, *construction*, or *silence*) to an urban soundscape recording at soundscape-to-masker ratios [32]. The maskers are AEs. Hence, ARAUS meets the needs of SoundAQnet training with affective supervision information. Unfortunately, ARAUS does not have AS and AE labels.

The USotW [49], as synthesis material for ARAUS, contains 360-degree video clips with GPS locations, which allows us to easily identify their corresponding scene. There are 3 classes of ASs, namely {*public square*, *park*, *street traffic*}. Following ARAUS synthesis rules, we obtain the AS labels of ARAUS.

Although six types of AEs have been explicitly added in ARAUS, we cannot directly use the six labels as AE labels because USotW already contains numerous AEs. To obtain the detailed AE labels in ARAUS, we first use the excellent pre-trained audio model PANNs [12] to label each audio clip with a one-second-level pseudo-label. Since the PANNs model is trained on AudioSet, a large-scale dataset with 527 classes of AEs, each one-second audio clip is assigned with a 527-dimensional soft pseudo label, corresponding to the probability of 527 classes of AEs within this second. Then, the soft pseudo-labels are mapped into hard pseudo-labels consisting of $\{0, 1\}$ by a threshold of 0.5. After accumulating and sorting the hard pseudo-labels for all one-second segments, we obtain the number of occurrences for the 527 classes of AEs in ARAUS, ranked from high to low. After considering the six types of AEs added in ARAUS, a total of 15 AE labels are obtained, which are {*Bird*, *Animal*, *Wind*, *Water*, *Natural sounds*, *Vehicle*, *Traffic*, *Sounds of things*, *Environment and background*, *Outside, rural or natural*, *Speech*, *Human sounds*, *Music*, *Noise*, *Silence*}. For training SoundAQnet, only clip-level AE labels are needed to distinguish whether the target AE is within the input clip. Hence, we again use PANNs to label the clip-level 527 AE probabilities for each audio clip. Then, the probabilities of 15 classes of target AEs are taken out and binarized into hard labels using a threshold of 0.1.

In the experiment of ARAUS [32], the validation set has 5040 samples, while the test set has only 48 samples. The test set may be too small to effectively evaluate the performance of our model. Thus, we randomly shuffled and re-divided the

ARAUS dataset. In proportion, 19152 30-second audio clips are randomly selected as the training set; 2520 and 3576 audio clips are chosen as the validation and test sets, respectively. To avoid intersections between the three sets, the number of audio clips used in this paper is 25248, rather than 25440.

B. Experimental Setup of Acoustic Model

Mel feature. The setting of log Mel features follows that of PANNs [12]. The 64 Mel bins are extracted by STFT with a Hamming window length of 32*ms* and a hop size of 10*ms*, resulting in Mel features having 3000 frames.

Loudness feature. Loudness features are extracted directly using the *ISO_532-1.exe* loudness program[1] recommended by ISO 532-1:2017 standard (Zwicker method) [20]. The input audio files are calibrated with a *".wav"* file containing the calibration signal, which is a sine wave at 1 kHz 60 dB. Then, the loudness features are calculated in frames of 2*ms*, resulting in Zwicker-loudness with 15000 frames. We upload the modified Python code and files to the ***homepage***.

Training settings. Adam optimizer is used to minimize the loss, with a learning rate 5e-4 and batch size 32. Since SoundAQnet contains a total of 12 tasks for classification and regression, referring to the settings in ARAUS [32], this paper monitors the *ISOP* loss on the validation set in early stopping. Starting from the 10th epoch, if the validation loss value of *ISOP* does not decrease within 10 epochs, training is stopped. The model is trained for a maximum of 100 epochs. The model is trained 10 times without a fixed seed to obtain the mean performance over the 10 runs. Accuracy (Acc) and threshold-free AUC [15] are used to evaluate ASC and AEC results. The mean squared error (MSE) is used to measure the regression results. The AS and AE labels that we annotated for ARAUS, code, and trained models are all available on the ***homepage***.

C. Comparisons to Other Models

TABLE I
COMPARISON OF DIFFERENT MODELS ON THE TEST SET (BATCH SIZE=32).

#	Model	Param (M)	Inference		ASC	AEC	AQ regression
			time(ms)	GPU(GB)	*Acc.*(%)	*AUC*	MSE Mean
1	AD_CNN [32]	0.52	5.6	1.79	89.63	0.84	1.128
2	Baseline_CNN	1.01	4.3	1.18	87.87	0.92	1.315
3	Hierarchical_CNN	1.01	4.6	1.18	89.82	0.89	1.293
4	MobileNetV2	2.26	5.5	1.94	89.67	0.92	1.145
5	YAMNet	3.21	4.9	2.16	88.84	0.90	1.199
6	CNN-Transformer	12.29	4.4	1.24	92.80	0.93	1.339
7	PANNs [12]	79.73	19.0	5.67	93.57	0.90	1.156
8	SoundAQnet	2.70	18.9	3.22	**95.31**	**0.94**	**1.054**

There are no other models similar to SoundAQnet for simultaneously modelling the AS, AE, and emotion-related AQ. Previous studies on AQ in soundscapes often use traditional linear regression to predict some AQ response values, while recent deep-learning-based studies only focus on a few specific AQs [7] [27] [29]. Therefore, we compare SoundAQnet with deep-learning models that perform well for auditory scene and event analysis, i.e., ASC and AEC tasks, as shown in Table I.

In Table I, #1 refers to the CNN used in the ARAUS paper [32]. CNN in #2 is the baseline for benchmarking the multiscale convolution-based SoundAQnet. It consists of 4 convolutional layers, each with 16, 32, 64, and 128 filters, and their corresponding kernel sizes of 3, 5, 7, and 9, respectively. After the convolutional layers, there are parallel ASC and AEC layers and regression layers for AQs. Hierarchical CNN in #3 aims to identify AS based on the predictions of AE, exploiting the implicit hierarchical relationship between AS and AEs [15]. Compared with #2, the ASC in #3 is better, but the AEC is affected by the hierarchical relationship. MobileNetV2 in #4 is a lightweight CNN that uses depthwise separable convolution to reduce the computational cost [50]. YAMNet in #5 is a CNN-based baseline for AEC from Google. Given the excellent performance of Transformers on audio tasks [15], #6 proposes CNN-Transformer, an encoder from Transformer [51] is added after the convolutional layer in Baseline CNN, to combine the spatial feature extraction capability of CNN with the temporal modelling capability of Transformer. Compared with #2, the introduction of Transformer encoder in #6 enhances the model's ability to discriminate acoustic scenes and events, and improves its classification results, but its overall result on 8D AQ regressions is not as good as those of the pure CNN in #4. The reason may be that, compared with Transformer encoder modelling AQs from the hidden layer features with the global perspective, CNN relies on a fixed-size convolutional kernel and performs better in learning the hidden layer features from different local perspectives, which is beneficial for modelling unique representations of each AQ.

Overall, the proposed SoundAQnet, which simultaneously models AS, AE, and human-perceived AQ, achieves the best results in ASC, AEC, and affect-related regression tasks with a similar number of parameters as MobileNetV2. More information on computational efficiency, such as training/inference speed, GPU, and hardware requirements, see the ***homepage***.

D. Ablation Studies of SoundAQnet

1) Ablation study on acoustic features

Tables II and III present the performance of SoundAQnet on ASC, AEC, ISO Pleasantness (*ISOP*) and ISO Eventfulness (*ISOE*) regression, and emotion-related AQ regression tasks when using different acoustic features, respectively. They are the mean results over 10 runs. When using single-class acoustic features, SoundAQnet retains only the corresponding branches, and the number of nodes in the graph-based fusion layer is reduced by half, i.e., $n = 4$.

TABLE II
MEAN PERFORMANCE OF SOUNDAQNET ON THE TEST DATASET (PART 1).

#	Acoustic feature		ASC	AEC	*ISOP*	*ISOE*	*pleasant*	*eventful*
	Mel	Loudness	*Acc.* (%)	*AUC*	*MSE*			
1	✗	✓	73.61	0.868	0.116	0.129	0.993	1.161
2	✓	✗	94.07	0.934	0.112	0.116	0.943	1.093
3	✓	✓	**95.31**	**0.941**	**0.106**	0.115	**0.899**	1.068

TABLE III
MEAN PERFORMANCE OF SOUNDAQNET ON THE TEST DATASET (PART 2).

#	Mel	Loud-ness	*chaotic*	*vibrant*	*uneventful*	*calm*	*annoying*	*monotonous*
			MSE					
1	✗	✓	1.187	1.067	1.237	1.105	1.191	1.234
2	✓	✗	1.098	0.975	1.165	1.043	1.105	1.167
3	✓	✓	**1.079**	0.979	1.168	**0.999**	**1.083**	1.159

The comparison of #1 and #2 in Table II shows that for the acoustic environment-related ASC and AEC, the Mel

[1]https://standards.iso.org/iso/532/-1/ed-1/en

TABLE IV
MEAN PERFORMANCE OF 10 RUNS OF SOUNDAQNET WITH CONVOLUTION BRANCHES OF DIFFERENT KERNEL SIZES ON THE TEST SET.

#	Kernel size				Sub-branch	Node	RFS	ASC	AEC	*pleasant*	*eventful*	*chaotic*	*vibrant*	*uneventful*	*calm*	*annoying*	*monotonous*
	3	5	7	9	{Mel; Loudness}	n	Time (s)	*Acc.* (%)	*AUC*	*MSE*							
1	✔				S_1: {(3, 3); (3, 1)}	2	0.76	93.67	0.913	0.919	1.071	1.088	0.987	1.166	1.013	1.110	1.170
2		✔			S_2: {(5, 5); (5, 1)}		1.44	93.73	0.917	**0.904**	1.056	1.071	0.980	**1.141**	**1.000**	**1.080**	1.161
3			✔		S_3: {(7, 7); (7, 1)}		2.12	**94.03**	**0.921**	0.910	1.050	1.067	0.969	1.145	1.005	1.092	**1.150**
4				✔	S_4: {(9, 9); (9, 1)}		2.80	93.91	0.920	0.916	**1.049**	**1.058**	**0.963**	1.150	1.006	1.091	1.151
5	✔	✔	✔	✔	$S_1 + S_2 + S_3 + S_4$	8	2.80	**95.31**	**0.941**	**0.899**	1.068	1.079	0.979	1.168	**0.999**	1.083	1.159

feature is more effective than the loudness feature, consistent with previous research [31]. The reason for this is easy to understand, following the notations in Section IV-A, Mel features with dimension (N, 64) include spectral information, which is essential for recognizing sounds, while this essential information is lost in over-compressed loudness features with dimension (N, 1). Compared with Mel-based #2, the fused feature in #3 performs better in the regression of *ISOP*, *pleasant*, *chaotic*, *calm*, and *annoying*, as well as in classifications of AS and AE. This indicates that introducing Zwicker-loudness explicitly can help Mel-based SoundAQnet on partial AQ regressions, and AS and AE classification. The effect of adding loudness is stronger for *ISOP* (the valence axis in Fig. 1), which is in line with expectations [2] [20]. Since *ISOP* and *ISOE* are linear combinations of the 8D AQs and thus do not offer additional insight, we will omit their results in the following experiments due to space limitations.

2) Ablation study on multiscale sub-branches

The duration of AEs may vary between a few tens of milliseconds, such as bird chirps, and several minutes, such as music. AQs along the valence dimension, *ISOP*, may be determined by short sounds that are identified as calming or annoying. AQs that are aligned with high arousal, high *ISOE*, such as *eventful*, *chaotic*, and *vibrant*, are related to multiple changes in the sound environment and require longer time windows to be identified. This ablation study aims to explore whether these expected relationships can be linked to the kernel sizes used in different branches.

Table IV shows the performance of SoundAQnet using features with different scales. The scale of features, i.e., the convolution receptive field size (RFS), is determined by the convolution kernel size. The performance of SoundAQnet with single-scale kernel branches is shown in #1-#4 of Table IV. For the convolution branch with a kernel size of 3, 5, 7, and 9, the RFS of each branch's last convolution layer relative to the input acoustic features is 76, 144, 212, and 280, respectively. The corresponding RFS in seconds is shown in Table IV. From #1 to #3, when the kernel size is increased from 3 to 7, that is, the RFS is increased from 0.76*s* to 2.12*s*, SoundAQnet's performance on ASC and AEC is improved, but continuing to increase the kernel size does not lead to higher accuracy.

In Table IV #1-#4, the emotion-related 8D AQ regression tasks achieve good results, except for #1. This indicates that the length of audio clips input to SoundAQnet needs to be greater than 0.76*s* to effectively capture human-perceived AQs. For #2 at the 1.44*s* RFS, SoundAQnet outperforms #1 in predicting AQs *pleasant*, *uneventful*, *calm*, and *annoying*, but a small decrease in performance is seen when increasing the kernel size further. For #4 at the 2.80*s* RFS, SoundAQnet outperforms options with smaller kernel sizes in predicting AQs *eventful*, *chaotic*, and *vibrant*. The results of #1-#4 suggest that SoundAQnet is time-window-aware, just as people may need different time scales to perceive different AQs. Finally, #5, which combines branches with different kernel sizes and RFS, performs better on ASC and AEC tasks while still performing reasonably well in most AQ regressions. Other combinations of S_1, S_2, S_3, and S_4 have been explored, with more details shown in the ***homepage***. In short, with the cooperation of small and large-size convolution kernels, SoundAQnet extracts multiscale features suitable for the target tasks, and captures acoustic environment information from multiple perspectives, thereby improving the results.

3) Ablation study on multiscale embedding fusion

The multiscale output is given as $\mathbf{M}$ = ($m3$, $m5$, $m7$, $m9$) for Mel branches and $\mathbf{L}$ = ($l3$, $l5$, $l7$, $l9$) for loudness branches. We can obtain the fusion result F by fusing these outputs, and then feed $\mathbf{F}$ into the residual embedding layer. Table V shows SoundAQnet's performance with different fusion methods. The mean and variance of the MSE of 8D AQ regressions are shown as an overall metric. For #1 in Table V, they are added; for #2, they are concatenated; for #3, the Hadamard product is used, where $\odot$ is the element-wise product. SoundAQnet performs similarly based on the fusion of #1-#3. #4-#6 adopt the scaled dot-product attention (Att), a key component in the widely used Transformer [51].

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax(\mathbf{Q}\mathbf{K}^{\mathbf{T}}/\sqrt{d_k})\mathbf{V} \quad (6)$$

where $\mathbf{V}$=$\mathbf{K}$, and d_k is $\mathbf{K}$'s dimension. In this case, the similarity between $\mathbf{Q}$ and $\mathbf{K}$ is exploited to adjust the information in $\mathbf{V}$, so a more informative $\mathbf{V}$ can lead to better results. In #4, $\mathbf{M}$ acts as $\mathbf{Q}$ and $\mathbf{L}$ acts as $\mathbf{K}$, Mel-based representations are used as a query to adjust loudness-based representations, and the output result mainly relies on $\mathbf{L}$. The operation of #5 is the opposite of #4, and #5 performs better than #4 on ASC and AEC tasks. The reason is similar to the Mel-only and loudness-only models in Table II. Notably, both #5 and #6, which use

TABLE V
MEAN PERFORMANCE OF SOUNDAQNET WITH DIFFERENT METHODS FOR FUSING MEL AND LOUDNESS-BASED SUB-BRANCHES ON THE TEST SET.

#	Fusion: ($\mathbf{M}$, $\mathbf{L}$)	ASC	AEC	AQ regression
		Acc. (%)	*AUC*	*MSE Mean*
1	Addition: $\mathbf{F} = \mathbf{M} + \mathbf{L}$	94.34±0.75	0.936±0.006	1.070±0.084
2	Concat: $\mathbf{F} = Concat(\mathbf{M}, \mathbf{L})$	94.47±0.57	0.934±0.005	1.068±0.089
3	Hadamard: $\mathbf{F} = \mathbf{M} \odot \mathbf{L}$	94.65±0.51	0.937±0.004	1.071±0.092
4	Q_Mel: $\mathbf{F} = Att(\mathbf{M}, \mathbf{L})$	88.85±2.96	0.865±0.008	1.059±0.083
5	Q_Loudness: $\mathbf{F} = Att(\mathbf{L}, \mathbf{M})$	94.54±0.99	0.884±0.013	1.040±0.082
6	Att_Q_M_Q_L	94.67±0.70	0.898±0.009	**1.038**±0.080
7	CLAP attention feature fusion	92.61±1.05	0.912±0.006	1.066±0.080
8	Graph-based	**95.31**±0.77	**0.941**±0.007	1.054±0.091

loudness as **Q** to modulate Mel-based representations, perform better in AQ regressions. #6, which concatenates $Att()$ in #4 and #5, shows the best result for AQ regressions. #7 uses the attention feature fusion (AFF) for variable-length audio in contrastive language-audio pretraining (CLAP) [52].

$$X_{fusion} = \alpha X_{global} + (1 - \alpha) X_{local} \tag{7}$$

where $\alpha = f_{AFF}(X_{global}, X_{local})$ is a factor obtained by AFF [52]. Unlike CLAP, which only uses Mel features, SoundAQnet uses two types of features. Thus, for CLAP-AFF-based SoundAQnet, Mel-based branches and loudness-based branches calculate the corresponding X_{fusion_Mel} and $X_{fusion\_loudness}$, respectively. Then, these two are concatenated and fed into a 1-layer multilayer perceptron (MLP) for fusion. Overall, the graph-based multiscale embedding fusion improves ASC and AEC performance, and shows competitive overall performance in regressions of human-perceived AQs. Source code for these fusions can be found in the ***homepage***.

4) The impact of network depth

SoundAQnet contains three layers of Conv2D blocks, each referring to VGG [38] and consisting of two convolution layers. Table VI presents the number of parameters (Param.), multiply-accumulate operations (MACs), and GPU memory required to train SoundAQnet with different depths.

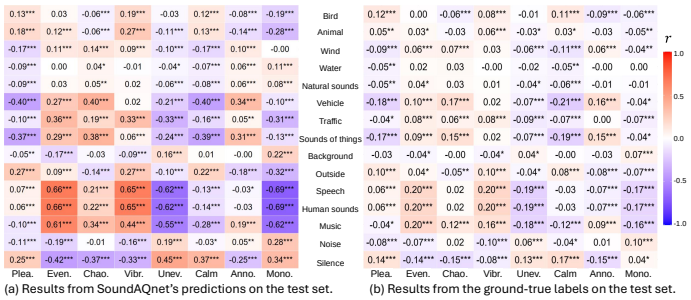
TABLE VI
SOUNDAQNET WITH DIFFERENT NUMBERS OF LAYERS. (BATCH SIZE=32)

# blocks	# layers	Para. (M)	MACs (G)	GPU (GB)	ASC *Acc. (%)*	AEC *AUC*	AQ regression *MSE Mean*
2	4	1.53	19.58	11.67	93.46	0.926	1.092
3	6	2.70	27.68	13.54	95.31	**0.941**	**1.054**
4	8	7.36	93.02	21.48	**95.60**	0.938	1.058
5	10	25.92	667.97	58.69	95.44	0.933	1.060

In Table VI, as SoundAQnet's depth increases, its computational overhead increases significantly but does not bring better results. SoundAQnet with the default number of blocks, three, balances model complexity, computational overhead, and achieves competitive performance in ASC, AEC, and AQ regression tasks. Due to space limitations, please also refer to the ***homepage*** for code and results for different dilation rates.

E. Correlation Study between AEs and AQs with SoundAQnet

People respond affectively and rate AQs based on recognition of various AEs, which are more related to AQ along the arousal axis [53]. This section answers the question whether SoundAQnet performs well in predicting AQs by implicitly learning the relationship between AEs and AQs. To investigate, Fig. 6 (a) shows the statistical significance of the predictions given by SoundAQnet on the test set of 3576 30-second binaural audio clips to analyze the relationship between AEs and the AQs they evoke. The Shapiro-Wilk test shows that the distributions of 15 AEs and 8D AQs are non-normal ($\alpha > 0.05$). Thus, we use Spearman's rho for correlation analysis between AEs and AQs. The statistical results in Fig. 6 (a) show that there are significant correlations between AEs and AQs. Specifically, some AEs like *'Bird'*, *'Animal'*, *'Outside, rural or natural' (Outside)* and *'Silence'* have significant positive correlations with pleasantness and calm. In addition, some AEs like *'Human sounds'*, *'Music'*, and *'Speech'* have significant positive correlations with eventful and vibrant, while some AEs, including *'Sound of things'* and *'Vehicle'*, can significantly evoke annoyingness (Anno.) and Chaotic. This indicates SoundAQnet's capability to capture the correlation between AEs and different AQs.



(a) Results from SoundAQnet's predictions on the test set. (b) Results from the ground-true labels on the test set.

Fig. 6. Spearman's rho correlation coefficients of AE and AQ. *, **, and *** denote statistical significance at the 0.05, 0.01, and 0.001 levels, respectively.

To further explore how SoundAQnet learns, Fig. 6 (b) shows the correlations on the ground-truth (GT) labels of the test set. This allows us to compare the differences in AE and AQ correlations between SoundAQnet predictions and the GT labels based on the same audio clips. Overall, the AE and AQ correlation trends in Fig. 6 (a) and (b) are consistent. However, the correlation trend in Fig. 6 (a) is stronger, indicating a more monotonous trend. SoundAQnet seems to accentuate correlations between specific AEs and AQs. For example, *'Animal'* correlates more significantly with all 8D AQs in Fig. 6 (a) than in (b). The stronger correlations in Fig. 6 (a) imply that SoundAQnet favours monotonous trends by reducing noise from the relationships it identifies as important.

F. Analyzing ASs and AQs with SoundAQnet

Fig. 7 shows the scatter plots, density, and marginal distributions of audio recordings of different scenes identified by SoundAQnet on the circumflex plane of affect (AQ plane). The density plots show that although there is a strong overlap between the distributions in the center of the plot, recordings of street traffic (ST) tend to be evaluated as more chaotic, recordings of parks tend to be rated as calmer, and recordings of public squares (PS) tend to be more vibrant. This difference between distributions is consistent with human intuitive expectations and matches previous research [54]. The marginal

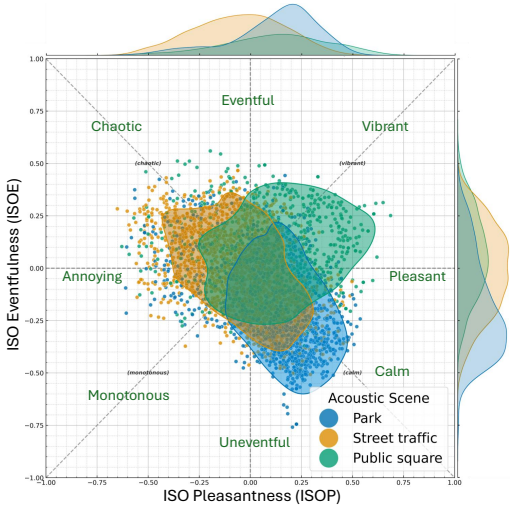


Fig. 7. The scatter plot, density distribution curves, and marginal distributions for the scenes involved are from SoundAQnet's predictions on the test set.

distributions of ISOP suggest that people may perceive similar pleasure in the park and PS scenes; the distribution of pleasure in the park is more concentrated than in PS, and people perceive less pleasure in ST than in the park and PS. The marginal distributions of ISOE suggest that the PS and ST scenes have similar event richness; ST's event diversity exceeds that of PS, and people perceive less event diversity in the park than in ST. Similar results were found based on human perception [55]. For more visualization plots, please see the ***homepage***.

G. SoundAQnet on the Cross-cultural Soundscape Dataset

The soundscape standard ISO/TS 12913-2 [35] gives researchers some freedom to incorporate cultural differences, but the basic idea behind the PAQ in [35] is still to try to remove cultural bias, starting with language. Soundscape attributes translation project (SATP) [56] aims at removing bias due to language and culture. The SATP dataset is a cross-cultural, multi-lingual, international, and cross-institutional dataset containing recordings with ISO-standard AQ labels. The SATP contains 19089 samples in 18 languages, consisting of 707 participants from 29 different national institutions evaluating 27 30-second binaural audio recordings according to ISO/TS 12913-2:2018 [35], in the institution's own language in each region. There are an average of 32 participants per institution.

To evaluate SoundAQnet's generalization ability across cultures, first, the Wilcoxon signed-rank test is used to compare the model's predictions on SATP with the ratings of 707 participants. The results show no significant difference between SoundAQnet's predictions and the average ratings of 707 participants ($p > 0.05$); that is, SoundAQnet performs similarly to the participants. Then, to explore the model performance in specific cultures, the consistency analysis of two cases in Europe and Asia is performed using Intraclass Correlation Coefficient (ICC) [57] analysis. As shown in Table VII, SoundAQnet's predictions show significant consistency with participants at University College London (UCL) on all AQs ($p < 0.05$); and with participants at Nanyang Technological University (NTU) on 7 AQs ($p < 0.05$) except Calm ($p > 0.05$). This indicates that SoundAQnet shows significant consistency on multiple affective dimensions across two different cultures, including Asian and Western cultures.

TABLE VII
ICC ANALYSIS FOR TWO DIFFERENT CULTURES (* IS $p < 0.05$); ICC RANGES FROM 0 (NO CONSISTENCY) TO 1 (PERFECT CONSISTENCY).

PAQ	Plea	Even	Chao	Vibr	Unev	Calm	Anno	Mono
UCL_ICC	0.89*	0.85*	0.89*	0.62*	0.82*	0.87*	0.82*	0.57*
NTU_ICC	0.78*	0.60*	0.90*	0.65*	0.61*	0.00	0.86*	0.53*

The results on SATP dataset show that SoundAQnet captures some degree of intra-cultural consistency. Prior studies show that LLMs generate different responses when queried in different languages, each embedding elements of its local culture [58]. Integrating SoundAQnet with LLMs may provide a good start for future studies aimed at fine-tuning the model to specific cultural contexts. In addition, we acknowledge that cultural differences in soundscapes are complex. Affective responses to soundscapes are inherently subjective and affected by personal, linguistic, and cultural factors, which is also one of the key challenges in soundscape studies.

VI. HUMAN EVALUATION (SOUNDSCAPER EXPERIMENT)

To assess the quality of captions generated by the SoundSCaper system, crowdsourced human evaluation is used to compare captions from SoundSCaper with captions annotated by two soundscape experts after cross-checking each other.

A. Experimental Design for Quality Assessment

The study employs a within-subjects design to evaluate soundscape captions from SoundSCaper and human experts. The sample size calculation is performed using G*Power [59]. The results of the calculation indicated that a sample size of 30 audio samples with $\alpha = 0.05$ and an assumed Effect Size of 0.5 for the Wilcoxon signed rank test achieved the pre-statistical power of 83.3%. Thus, the evaluation contains 60 audio clips from two distinct datasets. Dataset 1 (D1) contains 30 randomly selected samples with the same sound pressure levels (SPLs) from this paper's test set. Dataset 2 (D2) has 30 samples randomly selected from 5 external, i.e., model-never-seen, audio scene datasets, which are DCASE 2018/2019 [13], ISD [30], LITIS-Rouen [60] and a road sound dataset [61] of freeways with extreme noise environments commonly seen in daily life. The training set used in this paper contains 3 types of acoustic scenes, so recordings related to the 3 AS labels are selected from the 5 external datasets. Finally, the total duration of the D2 candidate data pool is about 1177 hours. The audio clips in D2 vary from 10 to 30 seconds with various SPLs without any limitations. Hence, D2 is used mainly to test the generalization performance of the SoundSCaper system.

1) Soundscape expert annotations: Two soundscape experts listen to randomly ordered samples and write captions in a style similar to the SoundSCaper caption example. This is done to ensure the consistency of caption styles generated by SoundSCaper and experts to prevent bias caused by participants guessing the caption's origin based on different styles.

2) SoundSCaper captions: As described in Section IV-B, SoundSCaper automatically generates target descriptions.

Finally, 120 soundscape captions are evaluated, 60 of which are derived from the proposed SoundSCaper system, and the remaining 60 are annotated by the two experienced soundscape experts. These captions are randomized. These captions were evaluated by a jury of 16 audio/soundscape experts and another jury of 16 laypersons, totaling 32 participants from 8 countries, including the UK, Singapore, Belgium, Canada, and France. Human assessment instructions, assessment data, and statistics' metadata are publicly available on the ***homepage***. Based on a self-assessment of the study's risks, ethical approval for this research was obtained from the Faculty of Engineering and Architecture of Ghent University.

B. Soundscape Caption Evaluation Metrics

Inspired by [62], we introduce the Transparent Human Benchmark for Soundscapes (THumBS) as a metric for the ASSC task. This indicator consists of precision, recall, and three other types of penalty items targeting specific defects.

1) Precision and recall $\in [1, 5]$: Precision (*P*) measures the accuracy of captions in describing the soundscape, specifically how well the caption's details match the actual sounds. Recall (*R*) evaluates the extent to which the caption captures the

comprehensive range of salient information (e.g., objects, attributes, relations) present in the soundscape.

2) Penalty items $\in [-2, 0]$*:* Fluency (*F*) assesses captions' textual quality as English prose, independent of its content accuracy. Conciseness (*C*) is used for repetitive descriptions. Irrelevance (*I*) is applied to the captions with details not present in the soundscape or unrelated to the sound content.

3) THumBS score: The final score can be calculated as

$$Score = (P + R)/2 + F + C + I \tag{8}$$

Due to limited space, we fully explain these metrics in the participant instructions presented on the ***homepage***.

C. Professionals' Evaluations

1) Comparison of SoundSCaper and expert captions

TABLE VIII
COMPARISON OF SOUNDSCAPE CAPTION QUALITY FROM SOUNDSCAPE EXPERT (E) AND SOUNDSCAPER (S) ON DATASETS D1 AND D2.

D		*precision*	*recall*	*fluency*	*conciseness*	*irrelevance*	Score
1	E	**3.84**±0.30	**3.93**±0.21	**-0.10**±0.07	**-0.14**±0.12	-0.22±0.12	**3.43**±0.35
	S	3.79±0.39	3.86±0.43	-0.12±0.09	-0.30±0.15	**-0.18**±0.15	3.22±0.53
2	E	**3.79**±0.39	**3.88**±0.43	**-0.15**±0.11	**-0.26**±0.12	**-0.26**±0.19	**3.16**±0.58
	S	3.64±0.39	3.64±0.29	-0.16±0.10	-0.27±0.16	-0.29±0.14	2.91±0.52

In the within-subject design study, the Shapiro-Wilk normality (SWN) test result shows that precision, recall and the final score data do not follow a normal distribution. Hence, we use the non-parametric Wilcoxon signed-rank (WSR) test. The results in Table VIII show that there is no significant difference between captions generated by SoundSCaper and those offered by soundscape experts on the final score ($p = 0.128$), and no significant difference between the two in terms of precision (p = 0.34) and recall (p = 0.44). This means that the quality of soundscape captions generated by SoundSCaper is comparable to that of soundscape expert-annotated captions. Fig. 8 details the precision, recall and final THumBS score in the evaluation of dataset D1. The horizontal line bisecting the box is the median, which coincides with the top line; the red dot represents the mean. The top and bottom borders of the box represent the 25th and 75th percentiles, respectively.

The intraclass correlation coefficient (ICC) [57] is used to assess the reliability of 16 raters' average ratings for 120 captions across 60 audio samples for Precision, Recall, Fluency, Conciseness, and Irrelevance. ICC values ($0.345 \sim 0.551$) show moderate to high agreement, with significant consistency ($p < 0.001$) and upper bounds of the confidence intervals above 0.6 for most criteria. These results indicate that 16 raters are sufficient to provide reliable assessments in this study.

For the mixed external dataset D2, the SWN test results indicate that the final score data on D2 do not follow a normal distribution ($p < 0.05$). Hence, we use a non-parametric WSR test. Table VIII shows that expert-annotated captions scored slightly higher than the captions generated by SoundSCaper; however, the WSR test shows that there is no significant difference between the two in the final scores ($p = 0.051$). As p is close to the significant level, we evaluate the ratings on the precision, recall, and penalty items, including fluency, conciseness, and irrelevance, separately. The distributions in Fig. 9 show that the precision and recall ratings follow a normal distribution, while the penalty items do not. Therefore, we use the paired t-test for precision and recall ratings; the result implies that there is no significant difference between the SoundSCaper-generated and expert-annotated captions on precision rating (p = 0.19) while there is a significant difference in recall rating (p = 0.028), which is not surprising as SoundAQnet is not trained on those datasets and the AE labels are also limited. The WSR test implies that there is no significant difference between SoundSCaper-generated and expert-annotated captions on fluency (p = 0.33), conciseness (p = 0.97), and irrelevance (p = 0.21). In summary, SoundSCaper has good generalization performance and adaptability, even though the recall rating of SoundSCaper-generated captions is significantly lower than that of the expert-annotated captions, and a competitive final score is still achieved.

2) Case study on the differences

The violin plot in Fig. 8 (c) reveals that the score distribution of SoundSCaper shows a slightly lower tail compared to that of the soundscape experts, indicating that SoundSCaper underperforms on some audio clips. Here, we explore the largest gap in final scores between SoundSCaper and human experts by subtracting the final score of SoundSCaper captions from that of soundscape expert annotations. The maximum value in the difference sequence is 1.99, observed in the sample *"28.flac"*. The soundscape captions are:

Human expert*: Immediately, there is a siren dominating the soundscape. As it fades away, it sounds like this is in a park, with people walking and chatting. Overall this is a calm soundscape, made somewhat annoying by the presence of the siren for part of the time.*

SoundSCaper*: In this park, you are surrounded by the soothing sounds of birds chirping, gentle human activities, and distant speech. The natural environment forms a tranquil backdrop. These sounds create a peaceful atmosphere, making you feel calm and content.*

In this case, the expert outperforms SoundSCaper, and emphasizes the disruptive presence of a siren in the park scene, highlighting its significant impact on the soundscape's calmness. Conversely, SoundSCaper paints a serene picture,

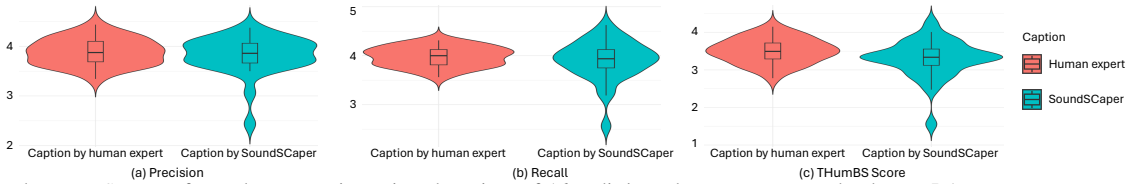


Fig. 8. *P*, *R* and *THumBS* score of soundscape captions given by a jury of 16 audio/soundscape experts on the dataset D1.

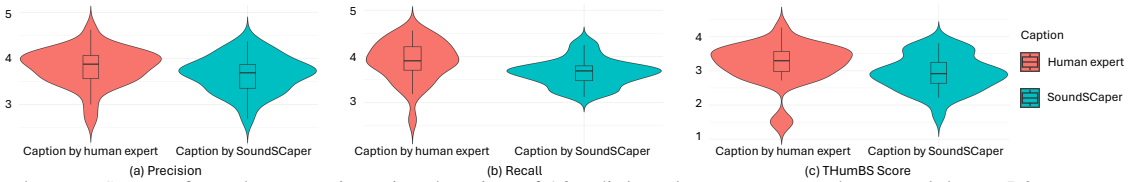


Fig. 9. *P*, *R* and *THumBS* score of soundscape captions given by a jury of 16 audio/soundscape experts on the external dataset D2.

not mentioning sirens and focusing only on peaceful elements such as birds chirping and distant speech. The reason is that there are no sirens in the 15 classes of AE labels in the dataset used for training the SoundAQnet, and hence it does not recognize this sound.

Next, we explore aspects where SoundSCaper outperforms human experts by subtracting the final score of expert-annotated captions from that of SoundSCaper captions. The maximum score difference is 0.84, the sample is *"26.flac"*, and its corresponding soundscape captions are:

Human soundscape expert: *This is a busy urban square with a mix of sounds. There is a constant hubbub of people talking and light music. There are also light vehicles passing regularly. The character is pleasant, lively, and comfortable.*

SoundSCaper: *In this bustling public square, music fills the air, accompanied by the chatter of people and sounds of things clinking and rustling. Occasionally, the rumble of vehicles and traffic noise can be heard in the background. The atmosphere is lively and vibrant, making the scene eventful and far from monotonous.*

In this case, the SoundSCaper caption is more appreciated by professionals. The expert's caption captures a mixture of pleasant, comfortable, and lively sounds in an urban square. SoundSCaper predominantly depicts the scene's lively and vibrant aspects, such as music and people's chatter, while downplaying vehicle noise. One possible explanation for this difference is that emotional feelings are subjective. People from various experiences and socio-cultural backgrounds may feel the same sound differently. Compared to descriptions with experts' individual responses to AQs, the AQ values predicted by SoundAQnet, trained on the ARAUS dataset of 25248 samples assessed by 605 participants, may be more consistent, less personalized, and more acceptable to other participants.

D. Layperson Evaluations

We employ the within-subject design for the layperson evaluation experiment. The SWN test shows that the final score data for D1 do not follow a normal distribution. Hence, we use the non-parametric WSR test. Table IX shows that there is no significant difference between captions generated by SoundSCaper and those provided by soundscape experts in terms of the final score ($p = 0.136$) by the layperson jury. That is, in the opinion of laypersons, the quality of captions generated by SoundSCaper is comparable to that of soundscape expert-annotated captions.

For the external dataset D2, the SWN test results show that the final score data do not follow a normal distribution, so the non-parametric WSR test is used. Table IX shows that in the evaluation by a layperson jury, SoundSCaper captions scored 3.24, slightly higher than the expert-annotated captions. However, the WSR test results show no significant difference in the final score between the two ($p = 0.90$).

TABLE IX
LAYPERSON EVALUATION OF SOUNDSCAPE CAPTION FROM SOUNDSCAPE EXPERT (E) AND SOUNDSCAPER (S) ON DATASETS D1 AND D2.

D		*precision*	*recall*	*fluency*	*conciseness*	*irrelevance*	Score
1	E	3.86±0.19	3.82±0.21	**-0.13**±0.08	**-0.10**±0.09	-0.18±0.12	**3.43**±0.33
	S	**3.90**±0.25	**3.84**±0.27	-0.15±0.10	-0.21±0.13	-0.18±0.10	3.33±0.36
2	E	**3.75**±0.38	**3.75**±0.35	-0.16±0.09	-0.18±0.12	-0.21±0.18	3.20±0.58
	S	3.72±0.40	3.67±0.36	**-0.12**±0.11	**-0.17**±0.15	**-0.17**±0.12	**3.24**±0.52

In addition, we compare the final scores given by 16 experts and 16 laypersons for both expert-annotated and SoundSCaper-generated captions. We use the paired t-test for expert ratings on both D1 and D2, as they follow the normal distribution. As for the final scores for SoundSCaper-generated captions, we employed the WSR test for both D1 and D2, as their ratings do not follow a normal distribution. The paired t-test results show no significant differences between the ratings of 16 experts and 16 laypersons for the expert-annotated captions on both D1 ($p = 0.32$) and D2 ($p = 0.54$). As for the final scores for SoundSCaper-generated captions, the WSR results show that there is no significant difference between the expert ratings and laypersons' ratings ($p = 0.58$) on D1; however, the laypersons' ratings for SoundSCaper-generated captions are significantly higher than the expert ratings on D2 ($p < 0.01$), laypersons rate the SoundSCaper-generated captions slightly higher than the expert annotations. This implies that laypersons are less sensitive to inconsistencies or subtle errors in soundscape captions compared to the experts, and appreciate the SoundSCaper-generated captions more.

E. Comparison with AAC systems

TABLE X
AVERAGE RESULTS OF AAC SYSTEMS AND SOUNDSCAPER ON VARIOUS NLP METRICS OF THE ASSC TASK ON DATASETS D1 AND D2, USING SOUNDSCAPE EXPERT CAPTIONS AS GROUND-TRUTH SENTENCES.

#	System	BLEU	ROUGE-L			METEOR	CIDEr
			Recall	Precision	F1 Score		
1	P-LocalAFT	0.0249	0.1001	**0.3428**	0.1537	0.0684	0.0029
2	ConvNeXt-Trans	0.1115	0.1454	0.2542	0.1831	0.1168	0.0917
3	GAMA	0.0879	0.1280	0.2852	0.1694	0.1009	0.0849
4	Qwen-Audio	0.1032	0.1479	0.2729	0.1868	0.1082	0.1213
5	SoundSCaper	**0.1901**	**0.2277**	0.2131	**0.2150**	**0.1610**	**0.2745**

In this section, SoundSCaper is compared with AAC systems with and without LLMs. Table X shows the performance of typical sequence-to-sequence (S2S) systems without LLM trained on the AudioCaps and Clotho-v2 datasets: P-LocalAFT [23] and ConvNeXt-Trans [22], and audio-LLMs such as GAMA [24] and Qwen-Audio [25]. The audio encoder of ConvNeXt-Trans is ConvNeXt pre-trained on AudioSet [47], and the decoder consists of a Transformer decoder with a structure similar to GPT. Given the excellent performance of ConvNeXt-Trans on AAC tasks [22], we fine-tune ConvNeXt-Trans on the training set of this paper as the baseline system for soundscape audio-to-text with affective qualities. GAMA [24] uses the audio spectrum transformer (AST), which consists of 12 Transformer encoder layers, for audio encoding, and uses multiple models such as multi-layer aggregators and Q-Former for enhancement, and then uses GPT-4 for caption generation. The number of parameters and the number of multiply-add-accumulate operations (MACs) required for AST in GAMA are 73.61MB and 108.99 G, respectively, while the corresponding SoundAQnet in SoundSCaper is only 2.70MB and 27.68 G. In terms of both the number of parameters and the model complexity, the proposed acoustic model, SoundAQnet in SoundSCaper, is much smaller than AST in GAMA. In short, Table X shows that SoundSCaper performs better than the LLM-free S2S AAC systems and the audio-LLMs on several NLP metrics. Please note that the soundscape expert captions have no influence on SoundSCaper's design, implementation, and training. In fact, they are collected after SoundSCaper is completed.

For source codes and data to fine-tune the audio-to-text baseline ConvNeXt-Trans for soundscapes, as well as scripts for NLP metrics, please see the ***homepage***.

F. Further Discussions

In the case study in Section VI-C, soundscape experts provided more specific and context-aware soundscape captions; e.g., they pointed out that the siren dominates the soundscape first, and after it fades away, people are chatting, which increased the spatiality and realism of the soundscape. In contrast, SoundSCaper mentions dominant AEs but lacks information about the corresponding sound sources and their spatial and temporal distribution. Due to the limited AS and AE labels in the used dataset, SoundAQnet cannot capture subtle but key sounds (such as short sirens) on the case *"28.flac"* as soundscape experts, resulting in descriptions lacking details. In addition, individuals from different cultural backgrounds have different AQs for soundscapes. The descriptions of the two soundscape experts from the UK and Sweden are based on personal experience, perceptions, and specialized training, and have a certain subjective style. Compared with the individual AQ responses of soundscape experts, the AQ values predicted by SoundAQnet, which is trained based on 25248 samples assessed by 605 participants, are more agreeable to other assessors in the case *"26.flac"*. This is consistent with SoundAQnet's cross-cultural performance on the SATP dataset, which contains 19089 samples in 18 languages and comprises 707 participants from 29 national institutions as discussed in Section V-G. In short, SoundSCaper's acoustic model, SoundAQnet, can effectively predict soundscape AQ in human evaluation based on datasets D1 and external mixed D2 and shows certain generalization abilities on the SATP dataset across countries and cultures. However, SoundSCaper still faces challenges posed by the inherent bias and subjectivity of individuals' affective response to sound and socio-cultural differences, which are key issues in soundscape research.

To assess the quality of ASSC, we investigated whether people would rate SoundScaper-generated and human-annotated captions differently after listening to soundscape recordings. In evaluations by 16 experts and 16 laypersons, captions produced by SoundSCaper are rated similarly to those produced as a consensus of two experts from different countries. The evaluation results show no significant difference in the rating behavior between the 16 experts and 16 laypersons. However, the 16 laypersons rated SoundSCaper higher on the D2 dataset, probably because laypersons are not as sensitive as experts to subtle differences and imprecision in soundscape captions.

In contrast to many AAC systems, SoundSCaper is not trained on datasets with captions. However, in Table X, where captions annotated by soundscape experts are used as ground truth, SoundSCaper outperforms AAC systems with and without LLM in NLP metrics. This shows that SoundSCaper outperforms general AAC systems in the ASSC task. However, since this study is the first attempt at ASSC, it faces some limitations: (1) the lack of datasets, preferably audiovisual datasets containing 360-degree videos and spatial audio, with high-quality soundscape captions produced by professionals; (2) the focus on common outdoor soundscapes, although the ARAUS dataset is based on USotW that contains recordings from 12 cities in 3 continents, it does not include extreme and rare soundscapes; (3) the subjectivity of individual perception of soundscapes and the differences in various regions and cultures need to be further explored. The proposed SoundSCaper's generalization performance in practical applications still needs to be optimized with more diverse data to ensure its adaptability to a wider range of soundscapes and user groups.

VII. Conclusion

Describing a soundscape, the acoustic environment as it is perceived and understood by people in context, is a cumbersome task. This paper formalizes the affective soundscape captioning (ASSC) task and designs the SoundSCaper system for this purpose. SoundSCaper used a novel lightweight acoustic model, SoundAQnet, to classify ASs and AEs and to predict AQs from soundscape, and also customized an LLM with prompt engineering techniques to turn such information into textual descriptions. Our work is significantly different from previous audio captioning studies, because it is the first attempt to jointly model the AS and AE in acoustic environments and the corresponding human-perceived AQ in soundscapes, and also the first attempt to connect the classification of AS and AE in acoustic environments with human affective descriptions of soundscapes. Comprehensive experiments and human evaluations have been performed to demonstrate the effectiveness of the proposed system for ASSC, as compared to relevant baselines. Overall, SoundSCaper offers competitive performance in human subjective evaluation and various objective captioning metrics, and the generated captions are comparable to those annotated by soundscape experts.

Given that this study is the first attempt at ASSC, it faces several limitations, such as the lack of available datasets with fully and carefully annotated soundscape recordings, and limited types of soundscapes and acoustic environments involved. In addition, AASC faces challenges posed by the inherent bias and subjectivity of individuals in their affective responses and socio-cultural differences, which are key issues in soundscape studies. Hence, the generalization of SoundSCaper needs to be further improved through diversified datasets and cross-scene and cultural validation. Introducing diverse large-scale data also helps the LLM in SoundSCaper fit AQs, explore scaling laws in AASC, and reduce the chance of LLM hallucinations, which remains a common issue. In addition, SoundAQnet performs comparably with human participants on the cross-country SATP dataset, implying that it has the potential to serve in unseen cross-cultural soundscape AQ assessments.

Acknowledgment

We appreciate the associate editor and all five reviewers for their insightful and helpful comments. We appreciate Dr. Francesco Aletta for valuable discussions and Dr. Gunnar Cerwen for professional soundscape captions. We thank 16 soundscape experts/professors in the human assessment, and another 16 general users. Due to limited space, we list their names on ***homepage*** to express our deep gratitude.

References

[1] M. Erfanian, A. Mitchell, et al., "Psychological well-being and demographic factors can mediate soundscape pleasantness and eventfulness," *Journal of Environmental Psychology*, vol. 77, pp. 101660, 2021.

[2] M. Erfanian, A. Mitchell, J. Kang, and F. Aletta, "The psychophysiological implications of soundscape," *International journal of environmental research and public health*, vol. 16, no. 19, pp. 3533, 2019.

[3] International Organization for Standardization, *ISO 12913-1:2014 Acoustics Soundscape Part 1: Definition and Conceptual Framework*.

[4] International Organization for Standardization, *ISO/TS 12913-3:2019 Acoustics Soundscape Part 3: Data Analysis*, 2019.

[5] Ö. Axelsson, M. E. Nilsson, and B. Berglund, "A principal components model of soundscape perception," *JASA*, vol. 12, no. 5, pp. 36–46, 2010.

[6] J. A. Russell, "A circumplex model of affect.," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161, 1980.

[7] Y. Hou, S. Song, et al., "Joint prediction of audio event and annoyance rating in an urban soundscape by hierarchical graph representation learning," in *Proc. of INTERSPEECH*, 2023, pp. 331–335.

[8] Y. Hou, Q. Ren, S. Song, Y. Song, W. Wang, et al., "Multi-level graph learning for audio event classification and human-perceived annoyance rating prediction," in *Proc. of ICASSP*, 2024, pp. 716–720.

[9] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational Analysis of Sound Scenes and Events*, Springer, 2018.

[10] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, and T. Virtanen, "Sound event detection in the dcase 2017 challenge," *IEEE/ACM TASLP*, vol. 27, no. 6, pp. 992–1006, 2019.

[11] Y. Hou, Q. Kong, S. Li, et al., "Sound event detection with sequentially labelled data based on connectionist temporal classification and unsupervised clustering," in *Proc. of ICASSP*, 2019, pp. 46–50.

[12] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM TASLP*, vol. 28, pp. 2880–2894, 2020.

[13] A. Mesaros, T. Heittola, and T. Virtanen, "Acoustic scene classification in DCASE 2019 challenge: Closed and open set classification and data mismatch setups," in *Proc. of DCASE*, 2019, pp. 164–168.

[14] Y. Hou, S. Song, C. Yu, W. Wang, and D. Botteldooren, "Audio event-relational graph representation learning for acoustic scene classification," *IEEE Signal Processing Letters*, vol. 30, pp. 1382–1386, 2023.

[15] Y. Hou, B. Kang, A. Mitchell, et al., "Cooperative scene-event modelling for acoustic scene classification," *IEEE/ACM TASLP*, pp. 1–13, 2023.

[16] K. Drossos, S. Adavanne, and T. Virtanen, "Automated audio captioning with recurrent neural networks," in *IEEE WASPAA*, 2017, pp. 374–378.

[17] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: an audio captioning dataset," in *Proc. of ICASSP*, 2020, pp. 736–740.

[18] OpenAI, "Chatgpt," https://chat.openai.com/, 2023, Accessed: 2024-4-1.

[19] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, et al., "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[20] International Organization for Standardization, *ISO 532-1:2017 Acoustics Methods for Calculating Loudness Part 1: Zwicker method*, 2017.

[21] A S. Koepke, A. Oncescu, J. Henriques, Z. Akata, and S. Albanie, "Audio retrieval with natural language queries: A benchmark study," *IEEE Transactions on Multimedia*, vol. 25, pp. 2675–2685, 2022.

[22] E. Labb, T. Pellegrini, and J. Pinquier, "CoNeTTE: An efficient audio captioning system leveraging multiple datasets with task embedding," *IEEE/ACM TASLP*, 2024.

[23] F. Xiao, J. Guan, H. Lan, Q. Zhu, and W. Wang, "Local information assisted attention-free decoder for audio captioning," *IEEE signal processing letters*, vol. 29, pp. 1604–1608, 2022.

[24] S. Ghosh, S. Kumar, A. Seth, C. K. Evuru, U. Tyagi, et al., "GAMA: A large audio-language model with advanced audio understanding and complex reasoning abilities," in *Proc. of EMNLP*, 2024, pp. 6288–6313.

[25] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, et al., "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023.

[26] L. Wang et al., "Assessing the affective quality of soundscape for individuals," *Science of Total Environment*, vol. 953, pp. 176083, 2024.

[27] P. Lundén and M. Hurtig, "On urban soundscape mapping: A computer can predict the outcome of soundscape assessments," in *Proc. of INTER-NOISE and NOISE-CON Congress*, 2016, vol. 253, pp. 2017–2024.

[28] M. G. Di Cesare et al., "Exploring the impact of soundscapes on emotional states: A data driven approach on environmental noise recording," in *E-Health and Bioengineering Conference*, 2024, pp. 1–4.

[29] M. Bradley and P. Lang, "Affective ratings of sounds and instruction manual, Technical report B-3," in *International Affective Digitized Sounds (2nd Edition IADS-2)*. 2007.

[30] A. Mitchell, T. Oberman, F. Aletta, et al., "The soundscape indices (SSID) protocol: A method for urban soundscape surveys with acoustical and contextual information," *Applied Sciences*, vol. 10, no. 7, 2020.

[31] Y. Hou, Q. Ren, H. Zhang, A. Mitchell, F. Aletta, et al., "AI-based soundscape analysis: Jointly identifying sound sources and predicting annoyancea," *JASA*, vol. 154, no. 5, pp. 3145–3157, 11 2023.

[32] K. Ooi, Z. Ong, et al., "Araus: A large-scale dataset and baseline models of affective responses to augmented urban soundscapes," *IEEE Transactions on Affective Computing*, vol. 15, no. 1, pp. 105–120, 2024.

[33] JL. B. Coelho, "Approaches to urban soundscape management, planning, and design," *Soundscape and the Built Environment*, pp. 197–214, 2016.

[34] Å. Skagerstrand, S. Stenfelt, S. Arlinger, and J. Wikström, "Sounds perceived as annoying by hearing-aid users in their daily soundscape," *International Journal of Audiology*, vol. 53, no. 4, pp. 259–269, 2014.

[35] International Organization Standardization, *ISO/TS 12913-2:2018 Acoustics Soundscape Part 2: Data Collection and Report Requirements*.

[36] P. Wang, P. Chen, Y. Yuan, D. Liu, et al., "Understanding convolution for semantic segmentation," in *IEEE WACV*, 2018, pp. 1451–1460.

[37] H. Gao, Z. Chen, and C. Li, "Hierarchical shrinkage multiscale network for hyperspectral image classification with hierarchical feature fusion," *IEEE JSTARS*, vol. 14, pp. 5760–5772, 2021.

[38] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[39] X. Bresson and T. Laurent, "Residual gated graph convnets," *arXiv preprint arXiv:1711.07553*, 2017.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of CVPR*, 2016, pp. 770–778.

[41] X. Lin, H. Chen, C. Pei, F. Sun, et al., "A pareto-efficient algorithm for multiple objective optimization in e-commerce recommendation," in *Proc. of ACM RecSys*, 2019, pp. 20–28.

[42] Z. Chen, V. Badrinarayanan, C. Lee, and A. Rabinovich, "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in *Proc. of ICML*, 2018, pp. 794–803.

[43] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. of CVPR*, 2018, pp. 7482–7491.

[44] C. Chen, R. Li, Y. Hu, S. M. Siniscalchi, P. Chen, E. Chng, and C. H. H. Yang, "It's never too late: Fusing acoustic information into large language models for automatic speech recognition," in *ICLR*, 2024.

[45] Z. Gekhman, G. Yona, et al., "Does fine-tuning llms on new knowledge encourage hallucinations?," *arXiv preprint arXiv:2405.05904*, 2024.

[46] Mi. Rumiantsau, A. Vertsel, I. Hrytsuk, et al., "Beyond fine-tuning: Effective strategies for mitigating hallucinations in large language models for data analytics," *arXiv preprint arXiv:2410.20024*, 2024.

[47] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, et al., "AudioSet: An ontology and human-labeled dataset for audio events," in *Proc. of ICASSP*, 2017, pp. 776–780.

[48] E. Fonseca, X. Favory, et al., "Fsd50k: an open dataset of human-labeled sound events," *IEEE/ACM TASLP*, vol. 30, pp. 829–852, 2021.

[49] B. De Coensel, K. Sun, et al., "Urban soundscapes of the world: Selection and reproduction of urban acoustic environments with soundscape in mind," in *Proc. of INTER-NOISE*, 2017, vol. 255, pp. 5407–5413.

[50] M. Sandler, A. Howard, M. Zhu, et al., "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. of CVPR*, 2018, pp. 4510–4520.

[51] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[52] Y. Wu, K. Chen, T. Zhang, Y. Hui, et al., "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proc. of ICASSP*, 2023, pp. 1–5.

[53] B. Schuller, S. Hantke, et al., "Automatic recognition of emotion evoked by general sound events," in *Proc. of ICASSP*, 2012, pp. 341–344.

[54] M. Carvalho, M. S. Engel, B. M. Fazenda, and W. J. Davies, "Evaluating the perceived affective qualities of urban soundscapes through audiovisual experiments," *Plos one*, vol. 19, no. 9, pp. 306, 2024.

[55] J. Y. Hong and J. Y. Jeon, "Influence of urban contexts on soundscape perceptions: A structural equation modeling approach," *Landscape and urban planning*, vol. 141, pp. 78–87, 2015.

[56] T. Oberman, A. Mitchell, F. Aletta, J. A. Almagro P., K. Jambrošić, and J. Kang, "Soundscape Attributes Translation Project (SATP) Dataset," Zenodo, Apr 2024, doi: 10.5281/zenodo.10993139.

[57] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: uses in assessing rater reliability.," *Psychological bulletin*, vol. 86, no. 2, pp. 420, 1979.

[58] M. Buyl, A. Rogiers, S. Noels, et al., "Large language models reflect the ideology of their creators," *arXiv preprint arXiv:2410.18417*, 2024.

[59] R. A. Mursa, C. Patterson, G. McErlean, and E. Halcomb, "How many is enough? justifying sample size in descriptive quantitative research," *Nurse Researcher*, vol. 3, no. 1, 2025.

[60] A. Rakotomamonjy and G. Gasso, "Histogram of gradients of time–frequency representations for audio scene classification," *IEEE/ACM TASLP*, vol. 23, no. 1, pp. 142–153, 2015.

[61] B. De Coensel, S. Vanwetswinkel, and D. Botteldooren, "Effects of natural sounds on the perception of road traffic noise," *JASA*, vol. 129, no. 4, pp. 48–53, 2011.

[62] J. Kasai, K. Sakaguchi, L. Dunag, J. Morrison, and R. L. Bras, "Transparent human evaluation for image captioning," *arXiv preprint arXiv:2111.08940*, 2021.