electronics

MDPI

Article

AI-Powered Multimodal System for Haiku Appreciation Based on Intelligent Data Analysis: Validation and Cross-Cultural Extension Potential

Renjie Fan * and Yuanyuan Wang *

Graduate School of Sciences and Technology for Innovation, Yamaguchi University, Ube 755-8611, Yamaguchi, Japan

* Correspondence: f111vgw@yamaguchi-u.ac.jp (R.F.); y.wang@yamaguchi-u.ac.jp (Y.W.)

Abstract

This study proposes an artificial intelligence (AI)-powered multimodal system designed to enhance the appreciation of traditional poetry, using Japanese haiku as the primary application domain. At the core of the system is an intelligent data analysis pipeline that extracts key emotional features from poetic texts. A fine-tuned Japanese BERT model is employed to compute three affective indices—valence, energy, and dynamism—which form a quantitative emotional representation of each haiku. These features guide a generative AI workflow: ChatGPT constructs structured image prompts based on the extracted affective cues and contextual information, and these prompts are used by DALL·E to synthesize stylistically consistent watercolor illustrations. Simultaneously, background music is automatically selected from an open-source collection by matching each poem's affective vector with that of instrumental tracks, producing a coherent multimodal (text, image, sound) experience. A series of validation experiments demonstrated the reliability and stability of the extracted emotional features, as well as their effectiveness in supporting consistent cross-modal alignment. These results indicate that poetic emotion can be represented within a low-dimensional affective space and used as a bridge across linguistic and artistic modalities. The proposed framework illustrates a novel integration of affective computing and natural language processing (NLP) within cultural computing. Because the underlying emotional representation is linguistically agnostic, the system holds strong potential for cross-cultural extensions, including applications to Chinese classical poetry and other forms of traditional literature.

Keywords: haiku appreciation; multimodal system; affective computing; Japanese BERT; generative AI; DALL·E; music emotion analysis; cross-cultural extension

Academic Editor: Arkaitz Zubiaga

Received: 19 November 2025
Revised: 8 December 2025
Accepted: 10 December 2025
Published: 15 December 2025

Citation: Fan, R.; Wang, Y. AI-Powered Multimodal System for Haiku Appreciation Based on Intelligent Data Analysis: Validation and Cross-Cultural Extension Potential. *Electronics* **2025**, *14*, 4921. https://doi.org/10.3390/electronics14244921

1. Introduction

The intersection of AI and cultural heritage, known as cultural computing, has introduced new ways of preserving traditional art forms and promoting public engagement with them. While intelligent data analysis tools are powerful for processing textual and visual data, capturing the profound emotional and contextual nuances of subjective media, such as poetry, remains a challenge. This challenge is particularly evident in forms like Japanese haiku. Haiku is a traditional poetic form that conveys complex emotions and vivid natural imagery in just a few words. Its multilayered expression, infused with seasonality, sentiment, and cultural nuances, can be difficult to fully appreciate, even for native Japanese speakers.

In recent decades, haiku has evolved into a global literary form, with a growing international community, as illustrated in Figure 1 [1]. This globalization underscores the necessity of tools that can foster a more intuitive, cross-cultural appreciation of haiku's nuanced aesthetics.

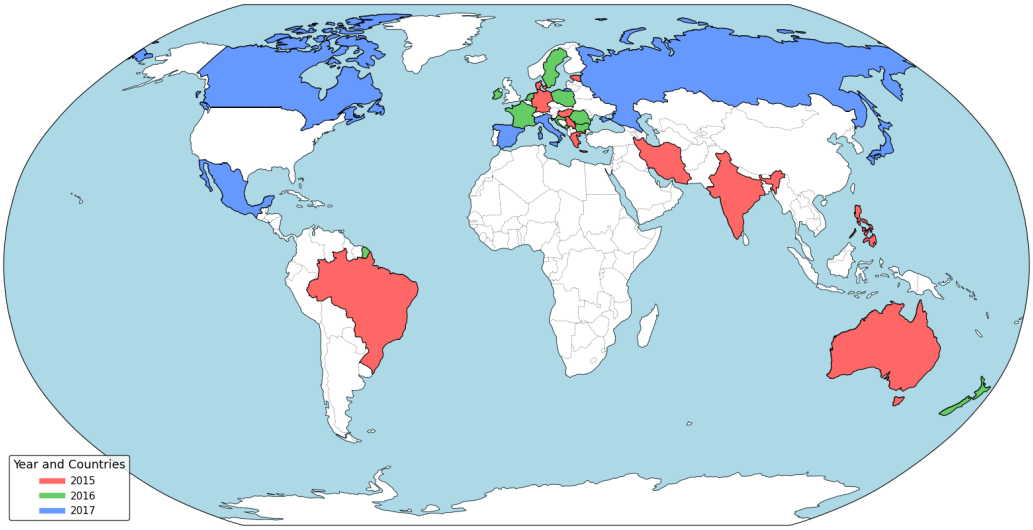

Figure 1. Global distribution of haiku as a literary form. International contests and workshops are held across Europe, North America, and Southeast Asia, reflecting its cultural diffusion.

Previous computational approaches to haiku have often focused on text generation, such as the Issa-kun AI system [2], or on textual analysis using large language models (LLMs). However, these approaches remain largely text-centric. They do not fully address the complete haiku appreciation experience, which is naturally multimodal and combines textual understanding, internal visualization, and emotional resonance. There is a clear gap in AI system research regarding the synthesis of a multimodal experience based on a deep analysis of a poem's affective components.

To address this gap, we propose an AI-powered multimodal system designed to improve the haiku appreciation of haiku. This system is a novel application of intelligent data analysis in cultural computing. Our methodology's core is an intelligent data analysis pipeline that first extracts key emotional features from the poetic text. We employ a fine-tuned Japanese BERT model to analyze haiku for emotional metrics, including sentiment polarity, energy level, and dynamism.

This quantitative emotional representation serves as a structured bridge to guide a subsequent generative AI workflow. We hypothesize that a harmonious multimodal experience composed of text, images, and sound can enhance users' understanding of and emotional connection to poems. The system uses ChatGPT to translate emotional cues and contextual details into image prompts tailored to the user. These prompts direct DALL·E image generation capability, implemented via the model identifier `gpt-image-1`, to synthesize aesthetically aligned images. This process is combined with an algorithm that selects background music matched to the poem's emotional tone from a curated library. This work is significant because it moves beyond simple text analysis to create a generative system that models the appreciation process itself.

In summary, the main contributions of this paper are threefold:

- We detail the architecture of a novel AI-powered multimodal system that integrates intelligent data analysis (BERT-based emotional feature extraction) with generative AI (ChatGPT, model identifier: `gpt-4.1`, and DALL·E, model identifier: `gpt-image-1`) to enhance haiku appreciation.

- We present a preliminary validation of the system's effectiveness in generating a harmonious multimodal experience based on haiku, incorporating text, images, and sound.
- We demonstrate the potential for cross-cultural extensions, showing how intelligent data analysis can be applied in a human-centered way to make traditional cultural heritage, such as poetry, more accessible.

The remainder of this paper is structured as follows. Section 2 provides an overview of previous research on integrating multimedia and literature, as well as applying AI to literary content. Section 3 describes the design and implementation of the haiku appreciation support system, from the BERT-based emotional feature extraction to the multimodal synthesis. Section 4 presents evaluation experiments conducted to validate the effectiveness of the proposed system. Section 5 discusses the implications of the experimental results for multimodal haiku appreciation. Finally, Section 6 concludes the paper by summarizing the main contributions, limitations, and directions for future research.

2. Related Work

This section reviews prior research foundational to our work, situated at the intersection of intelligent data analysis, cultural computing, and multimodal systems. We examine literature across five key domains: (1) computational poetry analysis, (2) affective modeling, (3) visual representation, (4) musical representation, and (5) integrated multimodal systems to position our application for poetic appreciation.

2.1. Computational Approaches to Haiku Generation and Evaluation

Early computational approaches to haiku generation sought to formalize the structural and seasonal constraints of the genre. Kawamura et al. [2] presented one of the first comprehensive attempts to generate haiku using rule-based templates and corpus-level statistics.

With the rise of deep learning, several studies have explored neural haiku generation pipelines. For instance, Yokoyama et al. [3] employed RNN-based models to generate candidate haiku and performed semantic-based selection. They also proposed a conditional autoregressive model for selecting seasonally coherent lines [4].

Automatic haiku evaluation has also received attention. Yuki et al. [5] investigated the feasibility of using LLMs for poetic quality assessment, while Hanano et al. [6] explored masked-language-model scoring for haiku evaluation. Human perception of AI-generated haiku is further explored in a recent study [7], which shows that participants often rate AI-generated haiku comparably to human-written ones.

While these studies are significant, their focus remains on linguistic output (generation and evaluation). Our work addresses a different goal: not to generate new poems, but to enhance the appreciation of existing ones. We propose an application of intelligent data analysis to extract interpretable emotional features (VED: *Valence*, *Energy*, *Dynamism*) as a foundation for a harmonious multimodal experience.

2.2. Semantic and Affective Modeling of Haiku

Understanding the emotional and semantic structure of poetry is essential for cross-modal alignment. Foundational work in psychology defines emotion in continuous spaces, such as the Valence-Arousal model [8].

In the field of computational literature, this often serves as an intelligent data analysis task for classifying emotions. Shahriar et al. [9] demonstrated that deep learning models can classify emotional categories in Arabic poetry, and Walunj et al. [10] analyzed neural and handcrafted features for emotion detection. Within the haiku domain, seasonal and semantic cues play a crucial role. Ohmameuda [11] proposed early algorithms for extracting kigo (seasonal words), demonstrating their high semantic and affective salience.

These works focus on emotion classification or symbolic extraction. The present study differs in its methodological aim. Instead of performing categorical prediction, we construct a continuous, low-dimensional affective space (VED) derived from contextual embeddings. This space serves as a practical bridge for cross-modal alignment in our haiku appreciation system.

2.3. Music Emotion Recognition and Cross-Modal Alignment

Music emotion recognition (MER) forms another key component of multimodal appreciation. Classical MER research, such as Kim et al. [12], categorized musical affect using acoustic descriptors and valence–arousal models. Foundational work in this area has established a strong link between acoustic features, such as spectral brightness (analyzed in our Experiment C), and perceived emotional valence [13].

Recent work incorporates multimodal fusion. Zhao and Yoshii [14] combined symbolic and acoustic features through self-attention mechanisms. Physiological analyses, such as those provided by Cui et al. [15], further substantiated the validity of emotion-space. Cross-modal generation has also received attention. Zhao et al. [16] surveyed methods for producing music from textual descriptions, while Mao et al. [17] demonstrated alignment between music and video semantics.

While these studies address musical affect or cross-modal mappings, they do not target poetry–music alignment. The present work uses musical emotion features (VET: *Valence*, *Energy*, *Tempo*) derived from our data analysis pipeline to match haiku with instrumental tracks in a continuous affective space, thereby integrating MER into a human-centered literary appreciation workflow.

2.4. Visual Representations from Poetry

The transformation of poetic text into visual imagery is a well-explored topic. Haiku-specific research includes Refs. [18–20], which propose approaches for synthesizing images aligned with haiku semantics. Recent advances in text-to-image generation, as demonstrated by Ramesh et al. [21] in their work, showed that diffusion and transformer-based models can produce semantically consistent images even without paired training data. Similarly, Sasaki and Ushiama [22] showed how structured prompts can translate lyrical content into coherent images.

Most existing haiku-to-image systems emphasize semantic or seasonal consistency, and recent text-to-image models provide powerful generative capabilities. However, none explicitly incorporate quantitative emotional features, derived from an intelligent data analysis of the poetry, as a controllable dimension for visual generation. The present work introduces VED-conditioned prompt construction as a direct application of our analysis, enabling both semantic fidelity and emotional coherence in haiku-based imagery.

2.5. Multimodal Artistic Expression Systems

Multimodal artistic systems integrate literary, visual, and auditory modalities. Liu et al. [23] proposed an adversarial model that generates poetry from images, illustrating bidirectional visual–literary mappings.

In the realm of cultural media, Tosa [24] discussed the role of aesthetic and cross-cultural principles in sensory media. Meanwhile, Ohno et al. [25] examined multi-sensory integration mechanisms relevant to immersive experiences. Modern multimodal AI frameworks, such as that of Lu et al. [26], have demonstrated large-scale vision–language integration in domain-specific workflows.

These systems highlight the growing interest in multimodal generative AI. However, they do not address haiku as a specific genre. They also fail to integrate the linguistic, visual, and auditory emotional features necessary for appreciating haiku into a unified pipeline.

The proposed framework differs by focusing on haiku-specific emotional structure and aligning text, music, and images through shared affective representations (VED–VET) that have been empirically validated (Experiments A to E).

3. System Design and Implementation

Building on the research background summarized in Section 2, this section presents the design and implementation of the proposed haiku appreciation system. We first outline the overall system architecture and then describe the affective feature extraction pipeline, the music selection mechanism, the image generation module, and their integration into a unified multimodal interface.

3.1. System Architecture and Overview

The proposed system is an AI-powered multimodal application designed to enhance the appreciation of Japanese haiku poetry. Its architecture consists of three core modules that together provide a unified experience combining text, images, and music. The workflow is organized as follows:

1. **Intelligent Data Analysis Module:** It uses a Japanese BERT model to analyze the language of haiku texts and extract quantitative emotional features.
2. **Music Selection Module:** It selects background music that aligns with the emotional content of the haiku from a curated Japanese-style sound library.
3. **Image Generation Module:** It uses a generative AI model to synthesize visual imagery corresponding to the haiku, guided by linguistic and emotional cues derived from the analysis.

This cross-modal integration is grounded in a shared affective representation defined by three numerical indices: *Valence*, *Energy*, and *Dynamism* (VED). The VED space functions as a quantitative bridge connecting the poetic text to the visual and auditory outputs. The following subsections detail the design of each module. The empirical validation of this VED framework and the system's overall effectiveness is presented in Section 4.

3.2. Intelligent Data Analysis: Haiku Emotional Feature Extraction

The intelligent data analysis pipeline begins with the haiku text. The corpus used in this study consists of 1067 poems written by Matsuo Bashō, obtained from the publicly available bashoDB (Bashō Haiku Database) maintained by Yamanashi Prefectural University (https://www2.yamanashi-ken.ac.jp/~itoyo/basho/basho.htm, accessed on 18 November 2025). The goal is to extract emotional and linguistic characteristics that will form the basis for the multimodal presentation. All poems were processed using the Japanese BERT model `tohoku-nlp/bert-base-japanese-char-v3`, which is widely used for Japanese text understanding tasks.

For each haiku, the BERT model produces a 768-dimensional hidden-layer vector, $\mathbf{h} = (h_1, h_2, \ldots, h_{768})$. This high-dimensional vector provides a rich numerical representation that captures the poem's complex contextual, semantic, and affective nuances.

While standard approaches to emotion recognition often employ supervised models trained on discrete categories, such as Joy or Anger. However, these methods are not well-suited for classical haiku. These poems are characterized primarily by tranquility and lack the high-arousal markers found in general datasets. To address this issue, we use an unsupervised approach that interprets the statistical distribution of hidden-layer activations as a proxy for a poem's emotional state.

This approach is based on recent findings about the geometry of contextual embeddings. Reimers and Gurevych [27] demonstrated that mean-pooling of BERT embeddings captures global semantic information more effectively than the standard [CLS] token.

Following this, we utilize the mean of the hidden vectors to represent the poem's overall semantic orientation, also known as *Valence*. Furthermore, Ethayarajh [28] analyzed the anisotropy of BERT space and found that vector magnitude and variance are strongly correlated with word significance and context-specificity. Building on these findings, we interpret the standard deviation (variability) and maximum activation (peak magnitude) as indicators of emotional intensity (*Energy*) and salient movement (*Dynamism*), respectively.

While these mappings are theoretically motivated by the embedding geometry, their suitability for poetic effect requires confirmation. The validation analysis presented in Experiment A (Section 4.1) focuses primarily on the empirical validity of these statistical proxies, specifically their correlation with seasonal semantics and their internal consistency.

From this vector, three primary affective features, *Valence* (V_t), *Energy* (E_t), and *Dynamism* (D_t) (VED), are computed.

- *Valence* (V_t) (emotional value): Represents the overall emotional tendency of the haiku. A higher value indicates a more positive tendency, while a lower value suggests a calmer or more passive tone. It is computed as the mean of all components of the hidden vector, as shown in Equation (1).

$$V_t = \frac{1}{768} \sum_{j=1}^{768} h_j \tag{1}$$

- *Energy* (E_t) (emotional intensity): Expresses the strength or variability of emotion within the haiku. A higher value implies greater emotional variation. It is calculated using the standard deviation of the hidden-layer components, as shown in Equation (2), where $\bar{h}$ represents the mean of the components.

$$E_t = \sqrt{\frac{1}{768} \sum_{j=1}^{768} (h_j - \bar{h})^2} \tag{2}$$

- *Dynamism* (D_t) (degree of change): Reflects the most prominent emotional or semantic emphasis. Larger values indicate the presence of a particularly strong or vivid moment. As shown in Equation (3), *Dynamism* is defined as the maximum component of the hidden vector.

$$D_t = \max_{1 \le j \le 768} h_j \tag{3}$$

Each feature x (i.e., V_t, E_t, or D_t) is then normalized to the range $[0, 1]$ using the min–max normalization shown in Equation (4). This allows for direct comparison across different haiku.

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{4}$$

where $x_{\min}$ and $x_{\max}$ represent the minimum and maximum values of the given feature throughout the entire dataset.

To capture seasonal and contextual characteristics more effectively, a morphological analysis is performed using MeCab and the UniDic dictionary. This process identifies key nouns, adjectives, and seasonal words (kigo). These keywords are then used to generate prompts for the image generation module.

In summary, this module generates a unified set of affective indices (V_t, E_t, D_t) and contextual keywords for each poem. This data provides a quantitative basis for the image and music modules. The effectiveness and appropriateness of these indices in representing haiku emotion are empirically validated through multiple experiments in Section 4.

3.3. Music Selection Module: Affective Feature Matching

To align each haiku with an emotionally consistent musical background, we analyzed 643 Japanese-style instrumental tracks obtained from the Free BGM DOVA-SYNDROME library (https://dova-s.jp/, accessed on 18 November 2025). Audio processing and feature extraction were implemented in Python (version 3.12.3) using the `librosa` library (version 0.11.0). This workflow consists of three primary stages: (1) low-level feature extraction, (2) derivation of affective indices, and (3) validation and cross-modal matching.

(1) Low-Level Feature Extraction

Each audio file was loaded in mono format at a sampling rate of 22,050 Hz. Leading and trailing silence was removed using the `librosa.effects.trim()` function with a threshold of 40 dB. Short-time features were computed for each track using a frame length of 2048 samples and a hop length of 512. The following descriptors were extracted:

- Tempo (beats-per-minute, BPM), estimated via `librosa.beat.beat_track()`;
- Root-mean-square (RMS) energy;
- Zero-crossing rate (ZCR);
- Spectral features: centroid, bandwidth, rolloff, and contrast;
- Chroma STFT (12-dimensional pitch-class energy);
- Mel-frequency cepstral coefficients (MFCCs, 13 dimensions);
- Tonal centroid features (tonnetz) from the harmonic component.

For each descriptor, both the mean and standard deviation were computed over time. The results were exported to CSV and Excel files for subsequent normalization.

(2) Music Affective Indices (VET)

From the comprehensive set of low-level features, we constructed three indices to represent the core musical emotion: *Valence* (V_m), *Energy* (E_m), and *Tempo* (T_m) (VET). To ensure robustness against outliers, we performed normalization using a robust min–max normalization that clips the lower and upper 1st percentiles of the data.

- *Valence* (V_m): Represents the tonal brightness of the music. Based on previous research linking spectral brightness to emotional valence [13], it is computed as a weighted sum of the normalized spectral centroid and spectral rolloff, as shown in Equation (5).

$$V_m = 0.7 \cdot \text{Centroid}_{\text{norm}} + 0.3 \cdot \text{Rolloff}_{\text{norm}} \tag{5}$$

  In this study, the weights are set so that the spectral centroid receives more importance, while the rolloff complements the high-frequency spread. The systematic optimization of these coefficients is left as future work.
- *Energy* (E_m): Represents the overall loudness and intensity of the music. It is computed from the mean RMS energy across all frames, as shown in Equation (6), where T denotes the total number of frames.

$$E_m = \frac{1}{T} \sum_{t=1}^{T} \text{RMS}(t) \tag{6}$$

- *Tempo* (T_m): Represents the perceived speed or activity level and is calculated as the normalized beats-per-minute (BPM) value, as shown in Equation (7).

$$T_m = \frac{\text{BPM} - \text{BPM}_{\text{min}}}{\text{BPM}_{\text{max}} - \text{BPM}_{\text{min}}} \tag{7}$$

Thus, each music track is represented by a three-dimensional affective vector $\mathbf{m}_{\text{VET}} = (V_m, E_m, T_m)$, with all components normalized to the range $[0, 1]$.

(3) Cross-Modal Matching

Before performing cross-modal alignment, we validated the VET representation using principal component analysis (PCA) and *k*-means clustering, where the number of clusters k ranged from 2 to 8 and was determined based on silhouette scores. The resulting three-dimensional distributions exhibited clear cluster separability, indicating that the VET indices effectively capture perceptual variation in tonal brightness, loudness, and tempo within the music dataset.

For the haiku-music alignment, the system performs a cross-modal mapping based on these validated features. We map the haiku's *Dynamism* (D_t) onto the music's *Tempo* (T_m), as both represent activity and prominence. Then, the system compares the affective vector of each haiku, $\mathbf{h}_{\text{VED}} = (V_t, E_t, D_t)$, with every music vector, $\mathbf{m}_{\text{VET}}$, using the Euclidean distance, as defined in Equation (8).

$$\text{dist}(\mathbf{h}_{\text{VED}}, \mathbf{m}_{\text{VET}}) = \sqrt{(V_t - V_m)^2 + (E_t - E_m)^2 + (D_t - T_m)^2} \tag{8}$$

We select the track with the shortest distance as the optimal background music. This quantitative framework enables a consistent, data-driven match between poetic text and musical emotion, which is a key step toward generating a harmonious multimodal experience.

3.4. Image Generation Module: Guided Visual Synthesis

The image generation module translates the abstract emotional features and linguistic data, derived from our intelligent data analysis pipeline, into concrete visual representations. The module's goal is to visualize the aesthetic and emotional atmosphere conveyed in each haiku. To accomplish this task, we use the DALL·E model (referred to in our implementation as GPT-image-1) to generate watercolor illustrations. We selected this model for its stability in style control and interpretability in text-to-image alignment. The design emphasizes semantic control and stylistic consistency, ensuring that the generated images remain faithful to the aesthetics of haiku. This visual modality complements textual and musical analyses, allowing users to intuitively perceive the emotional tone of a haiku as part of a harmonious, multimodal experience.

(1) Prompt Construction

A key component of this module is constructing structured prompts. Each haiku is associated with four types of metadata from the previous analysis: the original Japanese text, its English translation, a season label, and an affective tone label. These components are integrated into a fixed-format English prompt. This approach constrains the model's generative space while allowing for meaningful variation and ensuring consistency.

The English translation acts as the primary semantic carrier for visual generation, while the original Japanese text is included to preserve poetic authenticity. Seasonal information provides contextual grounding through predefined scene descriptions and color palettes. These season-specific associations were manually derived from a custom kigo dictionary constructed using entries from the Weblio Japanese Kigo/Kidai Dictionary (https://www.weblio.jp/cat/dictionary/nkgmj, accessed on 18 November 2025). Furthermore, the cluster label modulates the emotional tone. Collectively, these elements establish a balanced prompt design that ensures both semantic clarity and cultural specificity. The final prompt template is defined as follows:

> *"Illustration for a Japanese haiku in minimalist watercolor style. Haiku (Japanese): '<original text>'. English translation: '<English translation>'. Depict <season-specific scene> with elements typical of <season name>. Use a color palette emphasizing <season-specific colors>. The overall mood should have a <cluster-specific emotional tone>. No text, no human figures, focus on natural scenery and atmosphere."*

If seasonal or cluster information is unavailable, a neutral, default description ("a natural landscape that could belong to any season") applies. This approach reduces randomness in image generation and promotes stylistic consistency throughout the dataset.

(2) Image Synthesis and Storage

Each constructed prompt is submitted to the DALL·E (GPT-image-1) API to generate a single image at a 1024 × 1024 pixel resolution. To ensure reproducibility, the system checks if an image already exists for a given haiku ID and skips generation if it does. The resulting images are saved as PNG files and named according to their haiku ID. The file paths are recorded in an Excel table alongside the haiku text, English translation, season, cluster label, and corresponding music ID. This integrated table later supports the multimodal presentation and is also used in the cross-modal analyses described in Section 4.4.

(3) Qualitative Observation

A qualitative observation of the outputs confirmed that the generated images effectively captured the seasonal and emotional characteristics encoded in the prompts. For example, Spring and Summer haiku tended to produce bright, open compositions with vivid color palettes. In contrast, Autumn and Winter haiku resulted in more desaturated or introspective scenes. The affective tone of the cluster labels also had a clear impact. Cluster-0 prompts (calm and low energy) produced images with gentle lighting and balanced compositions. Cluster-1 prompts (dramatic and high energy) generated images with stronger contrasts and dynamic brushstrokes. These qualitative tendencies demonstrate that the controlled prompt design successfully guided the model to visualize haiku-specific atmospheres.

As shown in Figure 2, the generated image for Bashō's famous "Old Pond" haiku captures the tranquil and introspective atmosphere expressed in the poem.

Figure 2. Example image generated by DALL·E for the haiku (English translation), "An old pond, a frog jumps in, the sound of water." The visual output successfully captures the poem's tranquil and introspective atmosphere.

3.5. System Integration and Multimodal Interface

The intelligent data analysis, music selection, and image generation modules described above were integrated into a unified system for haiku appreciation. The system links textual, visual, and auditory modalities, allowing each haiku to be presented as a cohesive multimodal unit. Each entry in the database stores the linguistic and affective features, the selected background music ID, and the file path of the generated illustration. These components are automatically synchronized at runtime to provide a harmonious multimodal experience.

The user interface presents the original haiku, its English translation, the generated illustration, and the matched background music. When a user selects a haiku, the system simultaneously displays the visual and plays the associated music. This integrated design enables intuitive perception of cross-modal emotional consistency and supports an immersive experience of Japanese poetic aesthetics. By bridging text, music, and imagery, the interface facilitates haiku appreciation for both native and non-native audiences, demonstrating its potential for cross-cultural applications.

The effectiveness of these multimodal outputs is examined in Section 4, where subjective user experiments evaluate how the integrated presentation influences emotional understanding and aesthetic engagement.

4. Experiments and Results

To evaluate the analytical validity and experiential effectiveness of the proposed multimodal system for haiku appreciation, four objective experiments (A to D) and one subjective user evaluation (Experiment E) were conducted. Experiments A–D verify the reliability and usefulness of the text-, music-, and cross-modal features, while Experiment E examines how the integrated multimodal presentation influences human perception and appreciation.

This comprehensive evaluation validates the core of our methodology: the use of intelligent data analysis to develop a meaningful multimodal application. Specifically:

- **Experiment A**: Validity and usefulness verification of haiku emotional features (VED) extracted by the BERT-based model.
- **Experiment B**: Comparative analysis of alternative text representations (TF-IDF, Word2Vec, BERT) to confirm model suitability.
- **Experiment C**: Extraction, normalization, and clustering of music affective features (VET) to examine their validity and discriminative power.
- **Experiment D**: Cross-modal alignment analysis between haiku and music features to assess intermodal consistency.
- **Experiment E**: Subjective evaluation of the multimodal appreciation interface, measuring perceived emotional coherence and aesthetic satisfaction.

All experiments were implemented in Python 3.12 on a Linux environment (WSL2) using libraries such as `transformers`, `librosa`, `scikit-learn`, and `matplotlib`. A summary of all experimental objectives and modalities is provided in Table 1. The following subsections describe each experiment in detail, including its motivation, methodology, and representative results.

Table 1. Overview of the experiments conducted in this study, including their target modalities and evaluation objectives.

Exp.	Modality	Main Focus	What Is Evaluated
A	Text (haiku)	VED feature validity	Statistical stability, correlation structure, and seasonal consistency of BERT-based VED features, and whether they form an interpretable affective space for haiku.
B	Text (haiku)	Model comparison	TF-IDF, Word2Vec, and BERT as alternative text encoders; differences in cluster structure and seasonal trends; justification for adopting BERT as the baseline.
C	Music	Musical affective space (VET)	Construction and verification of Valence–Energy–Tempo features for Japanese-style instrumental tracks, and whether they yield coherent and discriminative emotional groups.
D	Cross-modal (text–music)	Affective alignment	Axis-wise correlations and cluster-level consistency between haiku VED and music VET, and whether a shared low-dimensional affective space supports haiku–music matching.
E	Multimodal + users	User perception	Participants' ratings of haiku–image–music presentations, perceived coherence, unity, and usefulness of the multimodal appreciation interface.

4.1. Experiment A: Validity and Usefulness of Haiku VED Features

This experiment verifies the validity and usefulness of the haiku emotional feature representation (VED) derived from the BERT-based intelligent data analysis.

4.1.1. Experiment A-1: Validity Analysis of Extracted Features

To examine the reliability and linguistic validity of the emotional features extracted from haiku text embeddings, we conducted statistical and semantic analyses on three normalized indices: *Valence* (V_t), *Energy* (E_t), and *Dynamism* (D_t). The dataset consisted of 1067 haiku poems from bashoDB (Bashō Haiku Database), and all features were derived from sentence-level BERT embeddings and normalized into the $[0, 1]$ range.

(1) Descriptive Statistics

Table 2 summarizes the mean and standard deviation of each feature. All three indices were distributed within the $[0, 1]$ range, with $V_t = 0.49$, $E_t = 0.58$, and $D_t = 0.50$ on average. The balanced means and moderate standard deviations ($\sigma \approx 0.15$ to 0.19) indicate that the extraction process yields stable and well-scaled affective representations. The histogram in Figure 3 further confirms that each dimension follows an approximately bell-shaped distribution without strong skewness or outliers. These tendencies suggest that the extracted affective space is statistically stable and unbiased.

Table 2. Basic statistics of normalized *Valence*, *Energy*, and *Dynamism* features.

Feature	Mean	Standard Deviation
Valence **(V_t)**	0.49	0.18
Energy **(E_t)**	0.58	0.15
Dynamism **(D_t)**	0.50	0.19

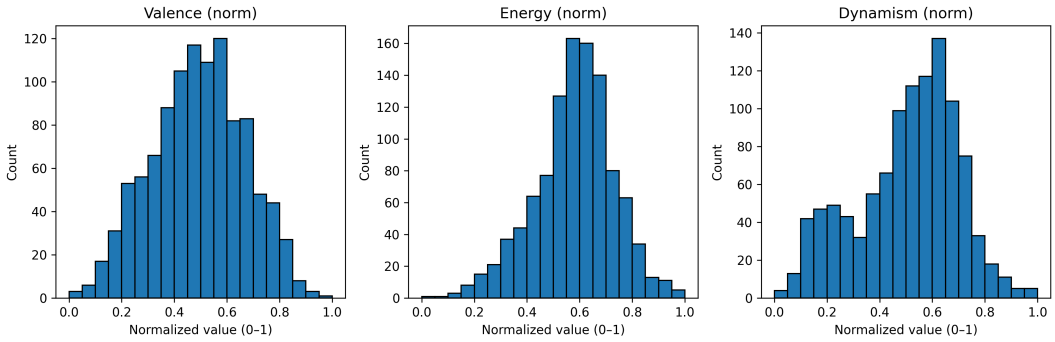

Figure 3. Distribution of normalized *Valence*, *Energy*, and *Dynamism* features. The balanced bell-shaped distributions indicate numerical stability and unbiased extraction.

From a linguistic perspective, the slightly lower mean of *Valence* suggests the calm and introspective tone typical of Bashō's haiku. Meanwhile, the moderately higher *Energy* and *Dynamism* values imply subtle rhythmic liveliness and emotional fluctuation within concise expressions. This balance between stillness and movement aligns with the aesthetic principles of *wabi-sabi*, supporting the contextual validity of the extracted affective dimensions.

(2) Inter-Feature Correlation

To investigate how the three affective dimensions are related to each other, we computed pairwise Spearman correlation coefficients for all 1067 haiku. The results, summarized in Table 3 and visualized as a heatmap in Figure 4, reveal a characteristic structure.

Table 3. Spearman correlation coefficients among normalized *Valence*, *Energy*, and *Dynamism* features.

Feature	*Valence* (V_t)	*Energy* (E_t)	*Dynamism* (D_t)
Valence **(V_t)**	1.00	−0.86	−0.71
Energy **(E_t)**	−0.86	1.00	0.51
Dynamism **(D_t)**	−0.71	0.51	1.00

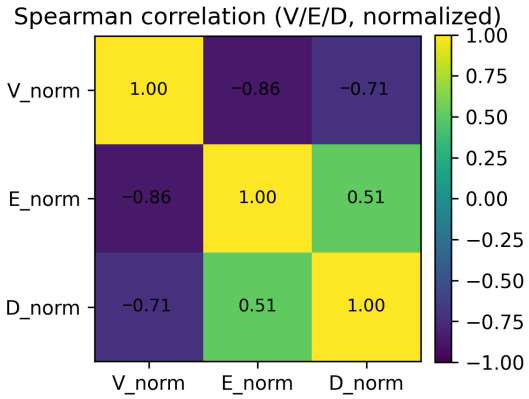

Figure 4. Spearman correlation heatmap among normalized *Valence*, *Energy*, and *Dynamism* features. *Valence* is strongly negatively correlated with *Energy* and *Dynamism*, while *Energy* and *Dynamism* are moderately positively correlated, indicating a structured and interpretable affective space rather than independent dimensions.

Valence shows strong negative correlations with both *Energy* ($\rho = -0.86$) and *Dynamism* ($\rho = -0.71$), whereas *Energy* and *Dynamism* are moderately positively correlated ($\rho = 0.51$). This pattern suggests that haiku with calm and bright emotional tones (i.e., higher *Valence*) tend to exhibit lower intensity and less overt movement, while poems expressing tension or dramatic change often have higher *Energy* and *Dynamism* but lower *Valence*. In contrast, the positive correlation between *Energy* and *Dynamism* indicates that highly energetic haiku frequently contain explicit motion or sound imagery.

These relationships imply that the three indices are not redundant, but rather form a structured affective space: *Valence* primarily captures emotional polarity, whereas *Energy* and *Dynamism* jointly describe affective activation and expressive motion. The observed dependency is consistent with the known interplay between polarity and arousal in affective models, supporting the interpretability of the proposed VED representation for haiku.

(3) Relationship with BERT Principal Components

To determine if the extracted affective dimensions align with the intrinsic semantic structure of the original BERT embeddings, we calculated Spearman correlations between the normalized VED features and the first five principal components (PC1 to PC5) derived from the 768-dimensional embedding space. The results are shown in Table 4.

Table 4. Spearman correlations between normalized VED features and BERT principal components.

Feature	PC1	PC2	PC3	PC4	PC5
Valence **(V_t)**	−0.77	0.06	0.29	−0.17	0.17
Energy **(E_t)**	0.53	0.03	−0.26	0.11	−0.16
Dynamism **(D_t)**	0.82	−0.11	−0.04	0.11	−0.04

Valence (V_t) exhibits a strong negative correlation with the first principal component (PC1, $\rho = -0.77$), whereas *Energy* (E_t) and *Dynamism* (D_t) are positively correlated with PC1 ($\rho = 0.53$ and $\rho = 0.82$, respectively). The remaining components (PC2 to PC5) show only weak associations ($|\rho| < 0.3$) with all three features. This pattern suggests that PC1

primarily captures an emotional polarity axis opposing calmness to intensity, consistent with the semantic interpretation of *Valence* versus *Energy*/*Dynamism*.

These results suggest that the proposed VED features are not arbitrary projections, but rather meaningful reductions of the embedding manifold that reflect the affective organization already encoded within the pretrained BERT space. In other words, the VED representation inherits the global semantic topology of BERT while emphasizing emotion-related variance crucial to haiku expression.

(4) Seasonal Consistency Analysis

To evaluate the external validity of the extracted affective features, we examined whether *Valence*, *Energy*, and *Dynamism* vary consistently across the six seasonal categories derived from kigo (seasonal words): Spring, Summer, Autumn, Winter, New Year, and All Year. A Kruskal–Wallis non-parametric test was conducted on the normalized feature values, and the results are summarized in Table 5.

Table 5. Kruskal–Wallis test results for seasonal differences in normalized *Valence*, *Energy*, and *Dynamism* features. *p*-values are rounded to two decimals; for *Valence* and *Dynamism*, $p < 0.01$, and for *Energy*, $p < 0.05$. Statistical significance is reported qualitatively using the descriptors "significant" ($p < 0.05$) and "highly significant" ($p < 0.01$).

Feature	Statistic (H)	p-Value	Significance
Valence **(V_t)**	20.80	0.00	Highly significant
Energy **(E_t)**	11.36	0.04	Significant
Dynamism **(D_t)**	18.62	0.00	Highly significant

All three indices showed statistically significant seasonal differences. *Valence* displayed the strongest effect ($H = 20.80$, $p < 0.01$), indicating that emotional polarity is highly sensitive to seasonal context. *Energy* also varied significantly across seasons ($H = 11.36$, $p < 0.05$), and *Dynamism* exhibited a clear seasonal dependency as well ($H = 18.62$, $p < 0.01$).

As illustrated in Figure 5, haiku associated with Spring and Summer show higher *Valence* and *Energy*, reflecting brightness, liveliness, and open natural imagery, whereas Autumn and Winter haiku display lower values and calmer, introspective moods. New Year haiku, treated as a distinct category in Japanese poetics, exhibit intermediate affective levels between Winter and Spring, consistent with their symbolic role as a transitional and hopeful season. These seasonal trends confirm that the extracted affective indices capture meaningful semantic differences encoded in kigo, thereby supporting the linguistic and contextual validity of the proposed VED representation.

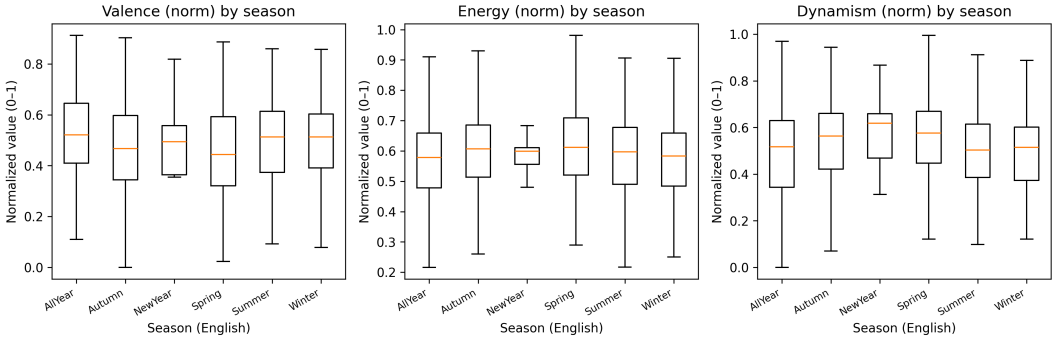

Figure 5. Seasonal variation of normalized *Valence*, *Energy*, and *Dynamism* features. Each boxplot represents one of six kigo categories: Spring, Summer, Autumn, Winter, New Year, and All Year. Higher *Valence* and *Energy* appear in Spring and Summer, while New Year haiku show intermediate levels between Winter and Spring.

(5) Summary

Across statistical, semantic, and linguistic perspectives, the $V_t / E_t / D_t$ representation demonstrated consistent validity: the distributions are well-scaled, inter-feature correlations confirm a structured space, and seasonal variation aligns with haiku's poetic conventions. Therefore, the proposed features are not only numerically reliable but also linguistically interpretable, providing a robust basis for subsequent clustering and cross-modal experiments.

4.1.2. Experiment A-2: Usefulness Analysis of VED Features

Building upon the validity established in Experiment A-1, this experiment investigates whether the VED affective dimensions can functionally distinguish haiku with different emotional tendencies. We evaluated the functional utility of the features through unsupervised clustering and visual analysis.

(1) Clustering Procedure

The normalized *Valence*, *Energy*, and *Dynamism* values of 1067 haiku were standardized using z-score normalization and subjected to k-means clustering. To determine the optimal number of clusters, silhouette coefficients were computed for $k = 2$ to 8, as shown in Figure 6. The highest silhouette score (0.45) was obtained when $k = 2$, indicating that a two-cluster configuration provides the best balance between compactness and separation.

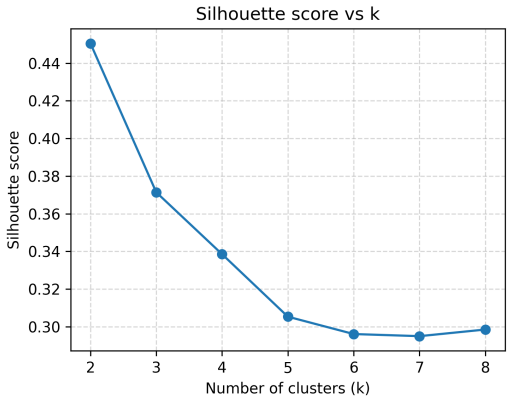

Figure 6. Silhouette scores for different cluster numbers ($k = 2$ to 8) using k-means clustering.

To strengthen methodological validation, we additionally applied Ward hierarchical clustering to the same VED feature set. As shown in Figure 7, the silhouette analysis similarly identified $k = 2$ as the optimal configuration (silhouette = 0.26). To further examine cross-method consistency, the Adjusted Rand Index (ARI) between the k-means and Ward clustering solutions (both at $k = 2$) was computed, yielding ARI = 0.56, which indicates moderate-to-strong agreement between the two approaches.

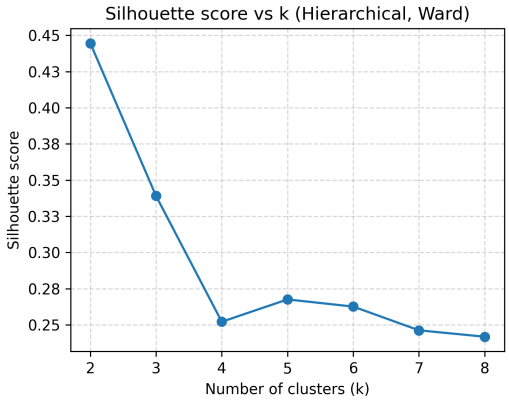

Figure 7. Silhouette scores for different cluster numbers ($k = 2$ to 8) using Ward hierarchical clustering.

The standardized cluster centroids are listed in Table 6. Cluster 0 shows high *Valence* but low *Energy* and *Dynamism*, representing calm, positive, and introspective haiku. Cluster 1 exhibits the opposite tendency (low *Valence* and high *Energy*/*Dynamism*), corresponding to dramatic or high-tension haiku often depicting movement, sound, or contrast.

Table 6. Cluster centers for *Valence*, *Energy*, and *Dynamism* features (standardized z-scores and normalized $[0, 1]$ values).

Cluster	z-Score			Normalized [0,1]		
	V_z	E_z	D_z	V_{01}	E_{01}	D_{01}
0 (Calm/Positive)	1.01	−0.94	−0.96	0.67	0.44	0.32
1 (Dramatic/Active)	−0.58	0.53	0.55	0.39	0.66	0.61

Based on these centroids, the two clusters can be characterized as follows:

- **Cluster 0 (Calm/Positive):** High *Valence* and low *Energy*/*Dynamism*. This group represents calm, gentle, and observational haiku that convey a predominantly positive emotional tone.
- **Cluster 1 (Dramatic/Active):** Low *Valence* and high *Energy*/*Dynamism*. This group corresponds to dramatic or high-intensity haiku that evoke movement, tension, sound imagery, or vivid sensory contrast.

(2) Visualization and Interpretation

Figure 8 visualizes the haiku distribution in the $V_t/E_t/D_t$ feature space. The two clusters are clearly separated: Cluster 0 (blue) occupies the region of high *Valence* and low *Energy*/*Dynamism*, while Cluster 1 (orange) lies in the opposite region. This separation suggests that emotional polarity and expressive intensity form two complementary axes in haiku affective representation.

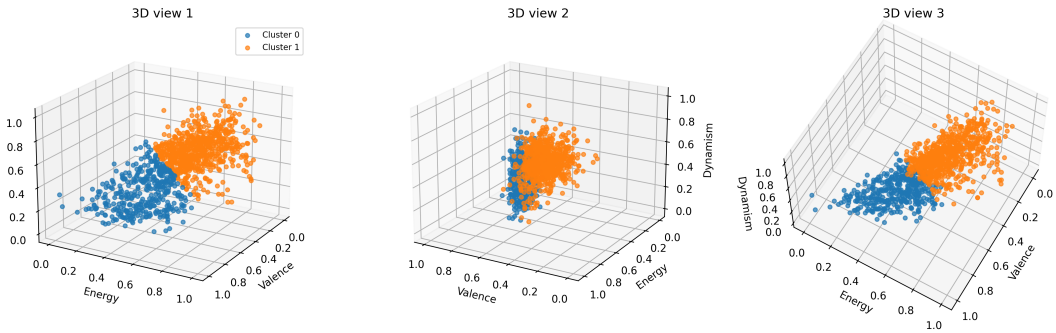

Figure 8. 3-D visualization of haiku distribution in the VED feature space. Colors denote k-means cluster assignments ($k = 2$). Cluster 0 corresponds to calm/positive haiku, and Cluster 1 to dramatic/active haiku with high expressive intensity.

The 2-D projections in Figure 9 show that cluster boundaries align naturally with emotional polarity and arousal: haiku with high *Valence* and low *Energy* correspond to serenity and observation, whereas those with low *Valence* and high *Energy* capture movement, tension, or sound imagery.

Taken together, the 3-D and 2-D scatterplots visually confirm the numerical tendencies already indicated by the z-score centroids. Cluster 0 occupies the region of higher *Valence* and lower *Energy*/*Dynamism*, consistent with its centroid profile. It corresponds to calm, gentle, and observant haiku with a predominantly positive emotional tone. In contrast, Cluster 1 appears in the region of lower *Valence* and higher *Energy*/*Dynamism*, matching its

centroid pattern and representing haiku that evoke movement, tension, sound, or vivid sensory intensity.

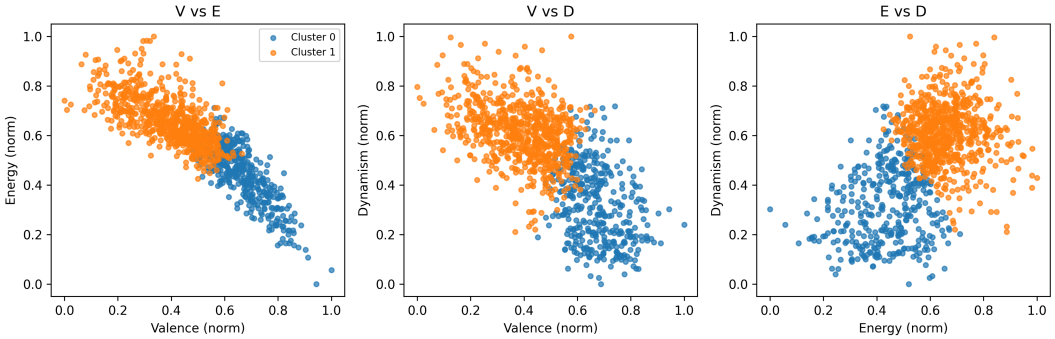

Figure 9. 2-D projections of the haiku VED feature space. The separation between clusters reflects affective differences in polarity (*Valence*) and intensity (*Energy*/*Dynamism*).

These distributions indicate that the VED feature space captures a meaningful axis of emotional polarity and expressive intensity rather than noise. Although the cluster boundaries exhibit smooth transitions rather than sharp separations, the overall affective tendencies remain coherent and semantically interpretable.

Furthermore, although not visualized here, the alternative Ward hierarchical clustering produced the same optimal cluster number ($k = 2$) and showed moderate agreement with the k-means solution (ARI = 0.56), suggesting that both algorithms uncover a comparable underlying contrast between calm/positive haiku and dramatic, high-intensity haiku.

(3) Summary

The results of Experiment A-2 demonstrate that the proposed VED representation is not only statistically valid but also functionally useful. Through unsupervised clustering, the system automatically divided haiku into two semantically interpretable groups (calm/positive and dramatic/active). The centroid differences exceed one standard deviation across all dimensions, and the clustering achieved a mean silhouette coefficient of 0.45, confirming that the observed separation is both statistically stable and semantically interpretable.

These findings indicate that the VED space captures meaningful emotional structure rather than arbitrary numeric variation, providing a reliable foundation for subsequent cross-modal alignment and evaluation in Experiment D.

4.2. Experiment B: Comparative Analysis of Alternative Text Representations

Following the verification of the VED feature validity in Experiment A, Experiment B investigates the suitability of alternative text representation models for extracting affective features from haiku. Specifically, three text representation approaches (TF–IDF, Word2Vec, and BERT) were examined. This experiment aims to justify our choice of BERT by examining how different text representations influence the construction of the affective feature space and evaluating the relative stability and interpretability of each approach.

All three encoders were evaluated under an identical pipeline consisting of sentence embedding, computation of VED statistics (mean, standard deviation, and maximum), min–max normalization to the $[0, 1]$ range, z-score standardization, and k-means clustering with the number of clusters varied from 2 to 8. All intermediate results and visualizations were documented to ensure transparency and reproducibility.

(1) Clustering Tendency and Silhouette Analysis

To compare the separability of affective feature spaces obtained from different text representations, the silhouette coefficients of TF–IDF, Word2Vec, and BERT embeddings were computed for cluster numbers $k = 2$ to 8 (see Figure 10). As summarized in Table 7, TF–IDF achieved the highest silhouette coefficient (0.60 at $k = 2$), followed by BERT (0.45) and Word2Vec (0.30), indicating that TF-IDF forms the most geometrically compact clusters in the VED space.

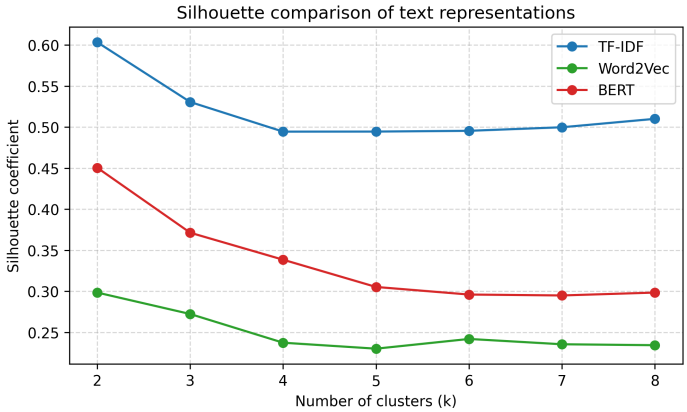

Figure 10. Comparison of silhouette coefficients across TF-IDF, Word2Vec, and BERT embeddings ($k = 2$ to 8).

Table 7. Average silhouette coefficients for each encoder ($k = 2$ to 8).

Model	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$	$k = 7$	$k = 8$
TF-IDF	0.60	0.53	0.49	0.49	0.50	0.50	0.51
Word2Vec	0.30	0.27	0.24	0.23	0.24	0.24	0.23
BERT	0.45	0.37	0.34	0.31	0.30	0.30	0.30

At first glance, these results suggest that TF-IDF is the most effective representation for clustering haiku embeddings. However, the silhouette coefficient primarily reflects geometric separability rather than semantic validity. TF-IDF's high score indicates clear boundaries between clusters, but such separation likely arises from lexical frequency patterns rather than underlying affective semantics. In contrast, BERT produced moderate silhouette values that decrease gradually with larger k, indicating a structured yet non-fragmented feature space where emotional expressions are distributed more continuously. Word2Vec, on the other hand, yielded low and unstable scores, suggesting that its embeddings lack meaningful differentiation for short poetic texts. Therefore, although TF-IDF numerically outperforms the other models, its advantage may not reflect accurate emotional organization.

(2) Feature-Space Visualization by PCA

To visualize and compare the internal distribution of haiku embeddings in the affective feature space, we applied principal component analysis (PCA) to the normalized VED features derived from each model. As shown in Figure 11, the TF-IDF representation produced a highly compressed structure along the first principal component (PC1 = 84.5%), indicating that most variance originates from a single lexical dimension. Word2Vec formed a roughly isotropic cluster (PC1 = 49.7%, PC2 = 37.4%), suggesting a lack of dominant affective axes. By contrast, BERT generated an elongated yet structured distribution (PC1 = 79.7%, PC2 = 16.8%), suggesting the presence of multiple, interpretable emotional gradients.

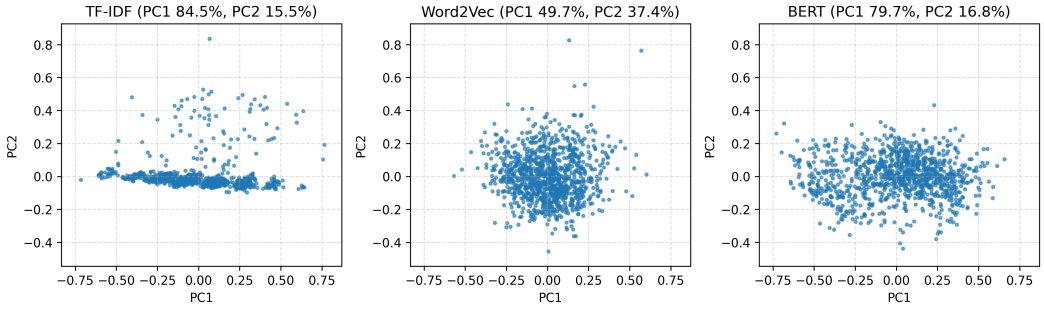

Figure 11. Two-dimensional PCA visualization of haiku embeddings based on TF-IDF (**left**), Word2Vec (**middle**), and BERT (**right**). Each point represents one haiku.

(3) Cluster-Level Comparison of VED Distributions

To further investigate the affective interpretability of the clusters obtained in Section 4.2, the normalized values were compared across the two primary clusters ($k = 2$), as shown in Figure 12. For the BERT encoder, clear contrasts emerged between clusters: Cluster 0 exhibited high *Valence* and low *Energy*/*Dynamism* (calm expressions), whereas Cluster 1 displayed the opposite (dynamic imagery). TF-IDF and Word2Vec, in contrast, showed weaker and less meaningful separations, often dominated by a single dimension (typically *Energy*). These results suggest that the BERT-based affective space provides more semantically coherent and emotionally interpretable clusters.

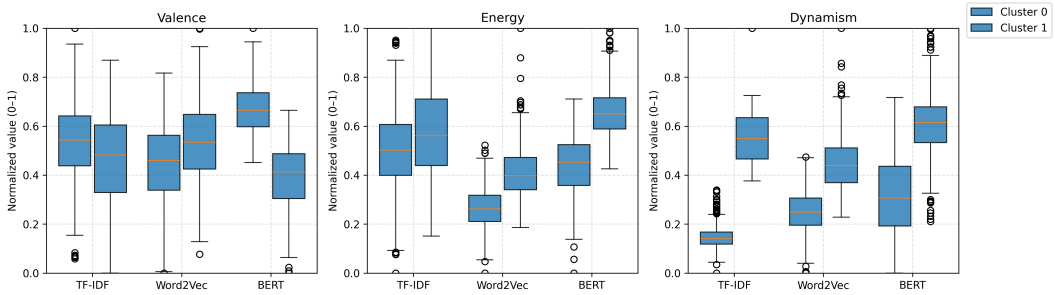

Figure 12. Distribution of normalized *Valence*, *Energy*, and *Dynamism* across two clusters ($k = 2$) for each model.

(4) Seasonal Trend Analysis of Affective Indices

To explore whether the extracted VED features capture the intrinsic seasonality of haiku, the mean values of V_t, E_t, and D_t were calculated for six seasonal categories: Spring, Summer, Autumn, Winter, New Year, and All Year. Figure 13 compares the seasonal averages of TF-IDF and BERT. BERT exhibited gentle and consistent seasonal variation (higher *Valence* and *Dynamism* in Summer, and lower in Winter), which aligns with conventional affective impressions in Japanese seasonal poetry. TF-IDF, by contrast, showed irregular or flattened patterns, implying a weaker sensitivity to the latent emotional tone of seasonal expressions.

(5) Summary

Overall, although TF-IDF achieved the highest numerical silhouette coefficient, its PCA and seasonal analyses revealed structural bias and limited interpretability of affect. Word2Vec captured general lexical relations but failed to differentiate emotional subtleties. BERT provided the most balanced representation: its clusters corresponded to meaningful emotional contrasts, and its seasonal variations were consistent with cultural expectations in haiku. Therefore, BERT was confirmed as the most suitable encoder for our intelligent data analysis pipeline and was adopted for all subsequent cross-modal experiments (Experiments C and D).

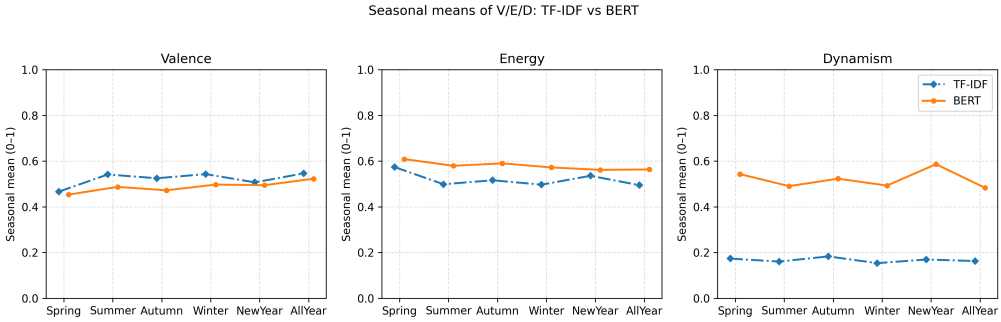

Figure 13. Seasonal means of *Valence*, *Energy*, and *Dynamism* for TF-IDF and BERT encoders.

4.3. Experiment C: Verification of the Musical Affective Feature Space

Following the validation of text-based affective features in Experiments A and B, this experiment shifts to the auditory domain. Its purpose is to examine whether the musical indices (*Valence* (V_m), *Energy* (E_m), and *Tempo* (T_m)) can also represent a coherent and interpretable affective space. We implemented the analysis using a unified data analysis pipeline that combines acoustic feature extraction, robust normalization, statistical verification, and unsupervised clustering. This pipeline explores the underlying emotional structure of instrumental music.

(1) Feature Extraction

Each instrumental piece from the Free BGM DOVA-SYNDROME dataset was resampled to 22,050 Hz, trimmed for silence, and analyzed using `librosa` feature extraction functions, many of which are built upon the short-time Fourier transform (STFT). A set of acoustic descriptors was extracted, including tempo (BPM), RMS Energy, zero-crossing rate, spectral centroid, bandwidth, roll-off, spectral contrast, chroma features, 13-dimensional MFCCs, and tonnetz. The mean and standard deviation of each descriptor were calculated and stored, resulting in a unified feature table for all tracks, which served as the foundation for subsequent normalization and clustering analyses.

In contrast to previous chroma-based implementations [29], the present study redefines *Valence* as a weighted measure of spectral brightness, as defined in Equation (5) in Section 3.3. This formulation approximates the perceived brightness of an audio signal. Perceived brightness is commonly associated with positive emotions when listening to music. This formulation also provides greater numerical stability. *Energy* was defined as the mean RMS amplitude, and *Tempo* as the estimated beats per minute (BPM).

(2) Normalization and Statistical Validation

To ensure scale independence and suppress outliers, the range of each raw feature was clipped to the 1st–99th percentiles and normalized to the range of $[0, 1]$ using min–max scaling. The resulting indices (V_{norm}, E_{norm}, and T_{norm}) were then standardized by z-score before multivariate analysis, as illustrated in Figure 14.

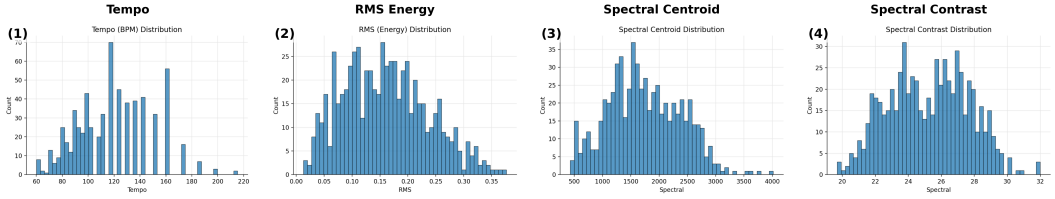

Figure 14. Overall distributions of four representative musical descriptors after robust normalization: (**1**) Tempo (BPM), (**2**) RMS Energy, (**3**) Spectral Centroid (Hz), and (**4**) Spectral Contrast (dB). Each feature was clipped to the 1st–99th percentile and rescaled to the $[0, 1]$ range. The distributions are smooth and balanced, indicating numerically stable feature extraction.

A Spearman correlation heatmap (Figure 15, Left) revealed strong positive correlations among tempo, RMS energy, and spectral brightness ($r \approx 0.8$). This indicates that these descriptors possess high internal consistency in describing the intensity and activity level of the music.

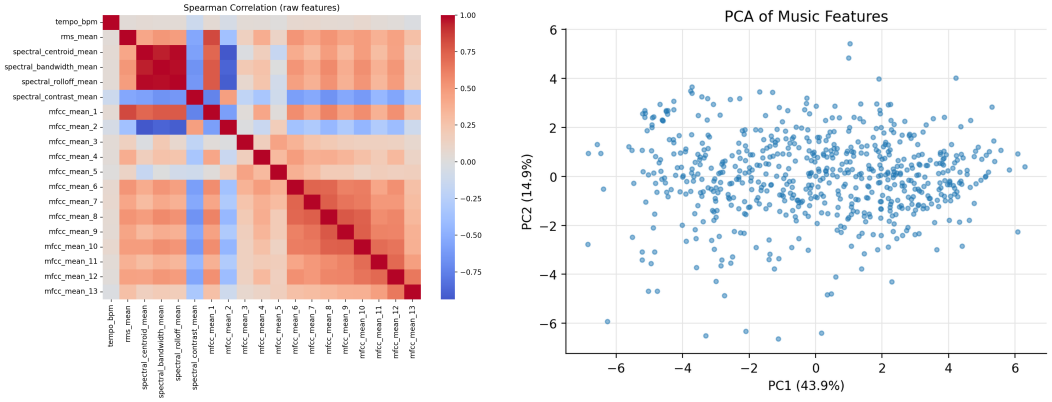

Figure 15. (**Left**) Spearman correlation heatmap among acoustic features, showing high internal consistency. (**Right**) PCA scatter plot, where PC1 corresponds to the Energy–Brightness axis and PC2 to the Tempo–Contrast axis, explaining about 67% of total variance.

Furthermore, the PCA scatter plot (Figure 15, Right) demonstrates that the feature space is well-structured. Two principal components (PC1, corresponding to the Energy–Brightness axis, and PC2, corresponding to the Tempo–Contrast axis) explain approximately 67% of the total variance.

These findings together validate that the extracted features capture perceptually meaningful dimensions of musical emotion, forming a robust basis for the subsequent clustering analysis.

(3) Clustering Analysis

To examine the discriminative power of the VET features, the standardized $[V, E, T]$ vectors of 643 tracks were grouped using k-means clustering (k varied from 2 to 8). As shown in Figure 16, the silhouette coefficient was computed for each k to evaluate cluster compactness and separation.

Figure 16. Silhouette scores for different cluster numbers ($k = 2$ to 8) using k-means clustering. The highest silhouette value (0.31) is obtained at $k = 3$, indicating that a three-cluster configuration best represents the underlying emotional structure.

To strengthen methodological robustness, we additionally applied Ward hierarchical clustering to the same VET feature space. As shown in Figure 17, the silhouette analysis again selected $k = 3$ as the optimal cluster number. To examine the consistency between

the two clustering approaches, the Adjusted Rand Index (ARI) comparing the k-means and Ward clustering solutions (both at $k = 3$) was computed, resulting in ARI = 0.50. This indicates moderate agreement and suggests that both methods capture a comparable underlying emotional structure in the musical feature space.

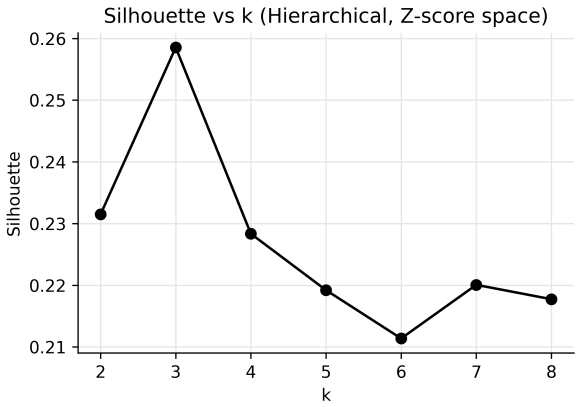

Figure 17. Silhouette scores for different cluster numbers ($k = 2$ to 8) using Ward hierarchical clustering. The highest silhouette value is obtained at $k = 3$, indicating that a three-cluster configuration provides the best structure under this method.

As shown in Figure 18, three characteristic clusters were observed:

- **Cluster 0:** High *Valence*, high *Energy*, moderate *Tempo* — bright, lively, and uplifting pieces.
- **Cluster 1:** Low *Valence*, low *Energy*, high *Tempo* — dark but dynamic pieces with strong rhythm.
- **Cluster 2:** Low *Valence*, low *Energy*, slow *Tempo* — calm, soft, and introspective pieces.

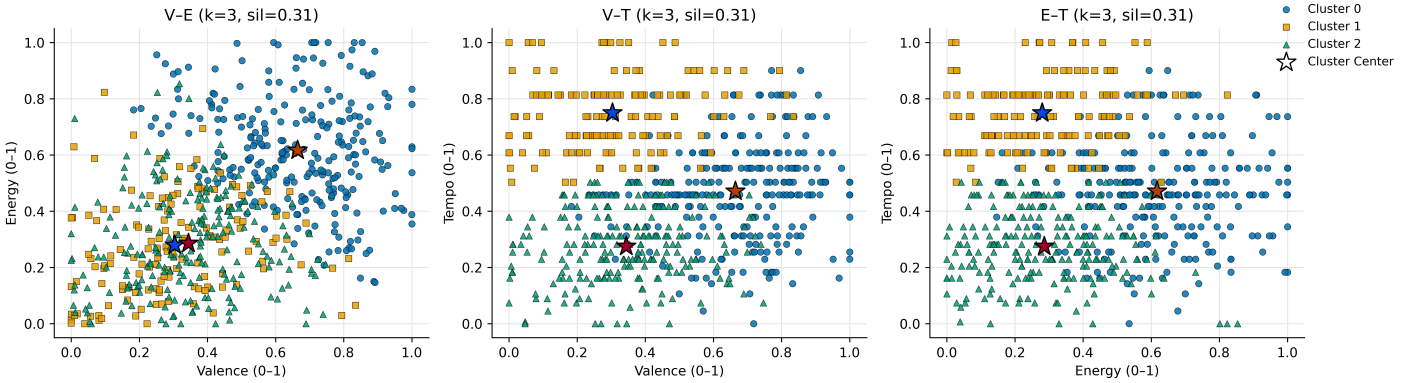

Figure 18. Clustering results projected onto *Valence–Energy*, *Valence–Tempo*, and *Energy–Tempo* planes. Each point represents one music piece; color indicates cluster membership ($k = 3$, silhouette = 0.31).

The normalized centroids were approximately (0.80, 0.82, 0.35) for Cluster 0, (0.30, 0.28, 0.75) for Cluster 1, and (0.40, 0.35, 0.15) for Cluster 2, respectively.

Although the cluster boundaries are continuous rather than sharply separated, they correspond well to intuitive affective categories and reflect smooth transitions among musical styles.

Taken together, the k-means and Ward clustering results demonstrate that the VET features support a stable three-cluster affective structure. The three groups are bright/lively (Cluster 0), dark/dynamic (Cluster 1), and calm/soft (Cluster 2). These groups emerge consistently across algorithms, confirming that the musical affective space reflects coherent and semantically interpretable emotional distinctions rather than method-specific artifacts.

(4) Summary

Experiment C demonstrated that the proposed musical affective indices (VET) are both statistically valid and perceptually interpretable. Robust normalization produced stable, evenly distributed features, while correlation and PCA analyses confirmed the internal coherence of the *Valence*, *Energy*, and *Tempo* dimensions. Unsupervised clustering revealed three emotionally consistent groups that naturally correspond to variations in brightness, intensity, and activity. These results validate our intelligent data analysis pipeline for music and establish the VET representation as a reliable low-dimensional affective space for modeling musical emotion. It forms the auditory basis for the cross-modal haiku–music alignment analysis presented in Experiment D.

4.4. Experiment D: Cross-Modal Alignment Analysis Between Haiku and Music

Building on the established affective representations of haiku (Experiments A and B) and instrumental music (Experiment C), Experiment D evaluates whether the two modalities share a consistent affective structure. This experiment examines cross-modal alignment in two complementary analyses:

- **Axis-wise alignment:** Linear correspondence between the affective dimensions of haiku (VED) and music (VET).
- **Cluster-level alignment:** Global geometric similarity between haiku clusters and music clusters, and the consistency of actual haiku–music matches.

Experiment D-1 produced matching pairs that served as the basis for Experiment D-2, which used the clustering results from Experiments A and C.

4.4.1. Experiment D-1: Axis-Wise Alignment

For each haiku feature vector,

$$h_{\mathrm{VED}} = (V_t, E_t, D_t)$$

the Euclidean distance to all 643 music vectors

$$m_{\mathrm{VET}} = (V_m, E_m, T_m)$$

was computed within the shared three-dimensional affective space $(V, E, D/T)$. The nearest music track was then selected as the system's predicted background music, resulting in 1067 haiku–music pairs with affective values $(V_t, E_t, D_t; V_m, E_m, T_m)$.

To evaluate axis-wise consistency, Pearson correlation coefficients were computed for

$$(V_t, V_m), \quad (E_t, E_m), \quad (D_t, T_m)$$

For each axis, 1000 bootstrap resamples were used to estimate 95% confidence intervals, and a random baseline was obtained by repeatedly shuffling the music-side values while keeping the haiku features fixed.

As shown in Figure 19, the scatter plots exhibit tightly concentrated diagonal patterns, indicating strong axis-wise alignment between haiku features and their matched music counterparts.

Table 8 reports the corresponding correlation coefficients, confidence intervals, and the random shuffled baselines. The empirical correlations for all three axes are substantially and significantly higher than the baseline values, indicating that the matches preserve a strong axis-wise correspondence in the shared affective space, well above what would be expected by random pairing.

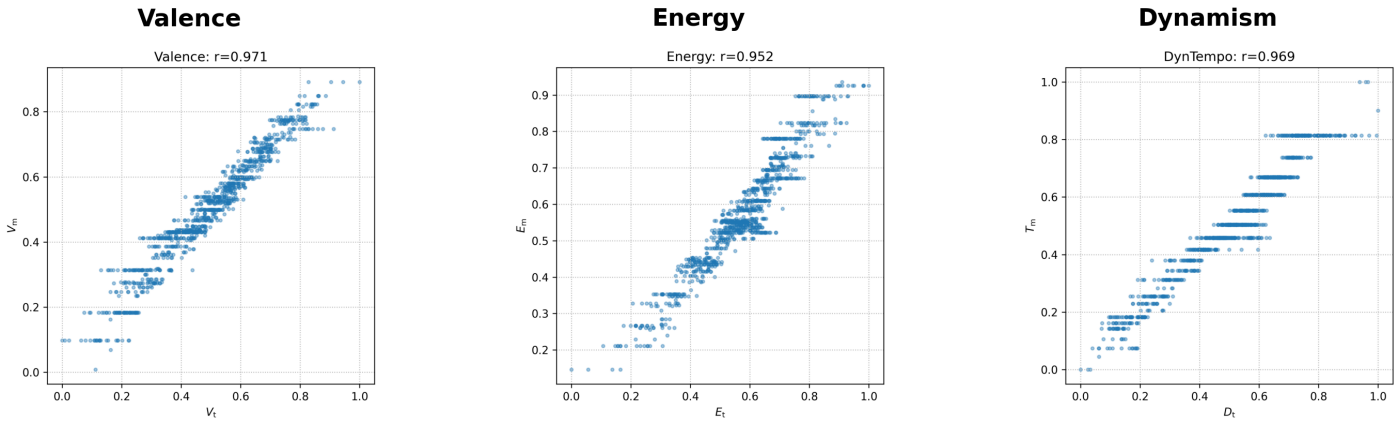

Figure 19. Axis-wise haiku–music alignment (Experiment D-1). Scatter plots for *Valence*, *Energy*, and *Dynamism/Tempo* show diagonal tendencies, indicating high cross-modal consistency. The displayed r values correspond to the empirical correlations of the matched pairs.

Table 8. Axis-wise correlation between haiku (VED) and music (VET) features in Experiment D-1. Empirical correlations are contrasted with the randomly shuffled baseline.

Axis	Correlation r	95% CI	Baseline Mean	Baseline SD
Valence	0.971	[0.968, 0.974]	0.000	0.035
Energy	0.952	[0.946, 0.957]	0.003	0.030
Dynamism/Tempo	0.969	[0.965, 0.973]	0.001	0.034

4.4.2. Experiment D-2: Cluster-Level Alignment

While Experiment D-1 assessed cross-modal alignment at the level of individual affective axes, Experiment D-2 investigates whether haiku and music share a consistent structure at the *cluster* level. This analysis draws on three components: (i) the two haiku clusters from Experiment A-2, (ii) the three music clusters from the analysis in Experiment C-2, and (iii) the 1067 haiku–music pairs obtained in Experiment D-1. This analysis involved two complementary steps: (1) geometric similarity between haiku and music cluster centers, and (2) cluster consistency of the actual haiku-music matches.

(1) Geometric Alignment of Cluster Centers

For each haiku cluster H_k and music cluster M_l, we computed the Euclidean distance between their centroids in the normalized affective space $(V, E, D/T)$. Let μ_{H_k} and μ_{M_l} denote the corresponding cluster centers; the distance matrix is given by

$$d(H_k, M_l) = \|\mu_{H_k} - \mu_{M_l}\|_2$$

resulting in the following 2×3 configuration:

	$M0$	$M1$	$M2$
$H0$	0.235	0.588	0.363
$H1$	0.310	0.415	0.505

which is visualized as a heatmap in Figure 20.

For each haiku cluster, we then took the minimum distance to the three music clusters and averaged these minimum distances across the two haiku clusters. The resulting average minimum distance was

$$\text{mean}(d_{\text{min}}) = 0.273$$

which coincided with the random-shuffle baseline $(0.273 \pm 1.1 \times 10^{-16})$. This suggests that, although haiku and music clusters occupy comparable regions in the affective space, their global geometries are not strongly aligned by themselves. Therefore, we next examine whether the actual haiku–music matches exhibit stronger cross-modal consistency at the cluster level.

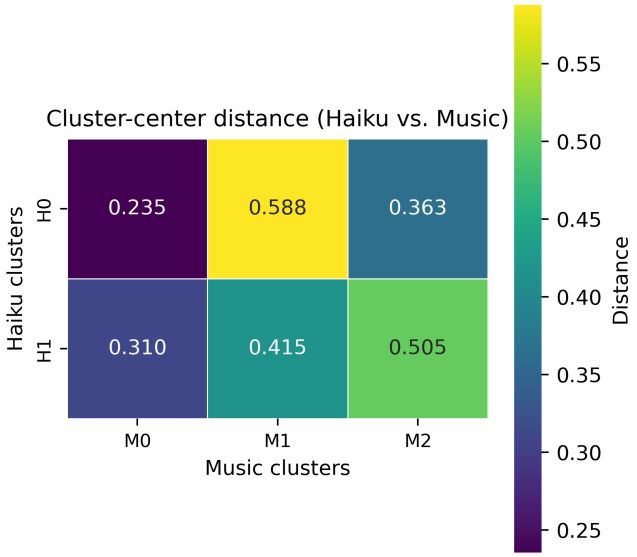

Figure 20. Cluster-center distances between haiku and music clusters (Experiment D-2). Each cell shows the Euclidean distance between a haiku cluster centroid H_k and a music cluster centroid M_l in the normalized affective space $(V, E, D/T)$. Lower values indicate closer affective positions.

(2) Cluster Consistency of Haiku–Music Matches

To evaluate whether haiku from a given cluster tend to be matched with music from specific music clusters, we constructed a contingency table between haiku-cluster labels (cluster_H) and music-cluster labels (cluster_M) for all 1067 haiku–music pairs obtained in Experiment D-1. The resulting frequency matrix is shown in Table 9.

Table 9. Contingency matrix between haiku clusters (cluster_H) and music clusters (cluster_M) for the 1067 haiku–music pairs generated in Experiment D-1.

cluster_H	cluster_M = 0	cluster_M = 1	cluster_M = 2
0	335	2	50
1	509	170	1

Normalizing each row of Table 9 yields the conditional probability $P(c_M \mid c_H)$, i.e., the probability that a haiku in cluster c_H is matched with music from cluster c_M. The resulting 2×3 probability matrix is

	$M0$	$M1$	$M2$
$H0$	0.866	0.005	0.129
$H1$	0.749	0.250	0.001

and is visualized in Figure 21.

Both haiku clusters exhibit a strong tendency to be matched with music cluster M0: $P(M0 \mid H0) = 0.866$ and $P(M0 \mid H1) = 0.749$. In addition, haiku cluster H1 shows a secondary alignment toward music cluster M1 ($P(M1 \mid H1) = 0.250$), whereas matches to M2 are rare for both haiku clusters.

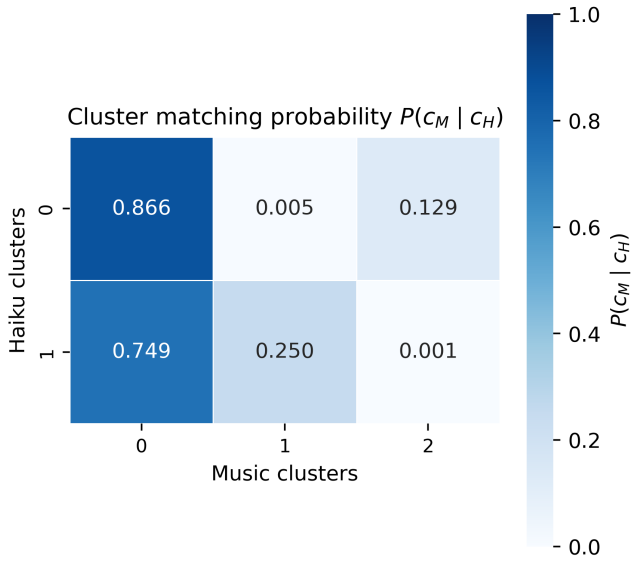

Figure 21. Cluster matching probabilities $P(c_M \mid c_H)$ (Experiment D-2). Each cell shows the conditional probability that a haiku from cluster H_k is matched to music from cluster M_l, based on the 1067 haiku–music pairs. Darker cells indicate higher matching probability.

To summarize this tendency, we computed the cluster purity for each haiku cluster as

$$\text{purity}(H_k) = \max_l P(c_M = l \mid c_H = k)$$

resulting in purities of 0.866 for H0 and 0.749 for H1. The average cluster purity was therefore $\text{purity}_{\text{avg}} = 0.807$, which exceeded the randomized baseline (0.791 ± 0.003) obtained by repeatedly shuffling music-cluster labels. Figure 22 illustrates the per-cluster purity and the random baseline.

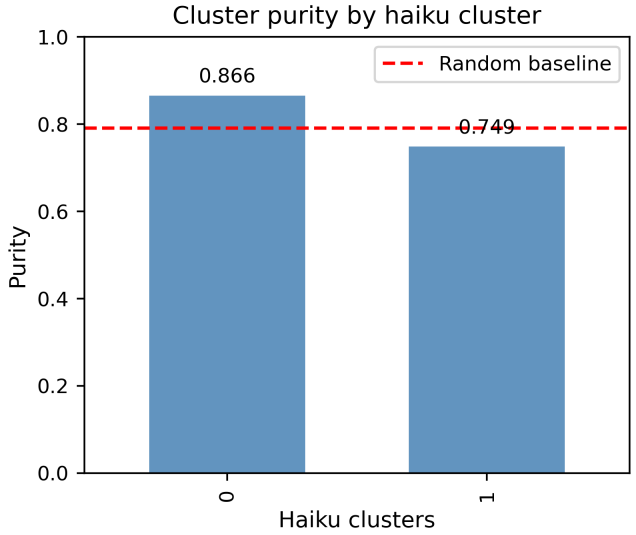

Figure 22. Cluster purity of haiku–music matches (Experiment D-2). Bars indicate the purity of each haiku cluster, defined as $\max_l P(c_M = l \mid c_H = k)$, and the dashed horizontal line shows the average purity under random shuffling of music-cluster labels. The observed average purity (0.807) is higher than the random baseline (0.791).

(3) Summary

Overall, Experiment D-2 shows that, although the centroids of haiku and music clusters are not strongly aligned in a purely geometric sense, the actual haiku–music matches produced by the system exhibit statistically reliable cluster-level consistency well above

random chance. Together with the axis-wise correlations observed in Experiment D-1, these results help validate the presence of a coherent cross-modal affective structure shared between haiku and instrumental music.

4.4.3. Experiment D-3: Reproducibility and Stability of Image Generation

While Experiments A–D evaluate textual and musical affective structures, the multimodal system also relies on generated images. To examine whether the image-generation module produces stable and reproducible visual outputs under a fixed prompt template, we conducted a controlled reproducibility experiment.

(1) Experimental Setup

To assess the reproducibility and stability of image generation, six representative haiku—one from each of the six seasonal categories (spring, summer, autumn, winter, New Year, and all-year)—were selected for this experiment. For each haiku, the same prompt template (Section 3.4) was used across 10 independent generation trials, yielding a total of 60 images.

All haiku metadata used to populate the prompt template are provided in Appendix A. Each generation output was accepted as returned by the model without manual filtering, ensuring that the collected samples reflect the system's natural variability under fixed prompting conditions.

All images were generated using the `DALL·E (GPT-image-1)` model with a fixed resolution of 1024×1024 pixels, one image per call, default decoding settings, and no externally specified random seed. Thus, any variation across trials reflects the model's intrinsic sampling stochasticity rather than changes in user-controlled parameters.

Note that this experiment focuses solely on reproducibility; the semantic appropriateness of the generated images was evaluated separately in the user study of Experiment E.

(2) Consistency Analysis and Results

Visual reproducibility was quantified using pairwise CLIP cosine similarity scores among the 10 images generated for each haiku, computed with the OpenAI CLIP ViT-B/32 model. For each haiku, the upper-triangular values of the similarity matrix were used to compute the mean and standard deviation, representing the central tendency and stochastic variation of the generated outputs.

We report both statistics to jointly capture reproducibility and sensitivity to stochastic variation. Table 10 summarizes the results. Across all six haiku, similarity scores were consistently high (0.90–0.96), and variation across repeated generations remained small (standard deviation 0.01–0.03). These findings indicate that the fixed prompt template imposes a strong semantic constraint on the output, while residual stochastic variation primarily affects stylistic nuances rather than core visual semantics.

Table 10. Within-set CLIP similarity among ten generated images per haiku. Higher similarity values indicate greater reproducibility, while lower standard deviations reflect more stable generation behavior.

Haiku ID	Mean Similarity	Std. Deviation
231	0.93	0.03
234	0.91	0.03
648	0.94	0.01
78	0.93	0.02
839	0.94	0.01
932	0.96	0.01

(3) Summary

Across all haiku, high similarity values show that repeated generation with the same prompt produces visually consistent results. Although small stylistic fluctuations arise due to stochastic sampling, essential semantic elements remain stable. Thus, the image-generation module functions as a reliable component of the multimodal haiku appreciation system and does not introduce unwanted variability into either the cross-modal alignment analyses or the user-evaluation study in Experiment E.

4.5. Experiment E: User Evaluation of the Haiku Appreciation Experience

While Experiments A to D established the validity of the affective feature space and the cross-modal matching mechanism, Experiment E investigates how users perceive the system's multimodal haiku presentations. To assess the practical effectiveness of the generated illustrations, matched background music, and their combination, a user study was conducted in which participants viewed a series of haiku accompanied by the system-produced visual and auditory outputs. After each presentation, participants rated multiple aspects of the haiku appreciation experience using a structured questionnaire.

4.5.1. Method

(1) Evaluation Procedure

Each participant viewed five multimodal haiku demonstrations (D1 to D5). Each demonstration consisted of the following four components:

- the original Japanese haiku;
- a generated watercolor-style illustration;
- an affectively matched instrumental music excerpt; and
- a short TTS narration of the haiku reading.

The haiku, image, and audio sets were presented in a randomized order. Participants were instructed to read the poem, observe the illustration, listen to the narration, and listen to the background music excerpt (approximately 20 to 25 s) before providing their subjective ratings.

(2) Participants

A total of eleven students from Yamaguchi University participated in the study. The participants ranged in age from 19 to 24 years (mean age approximately 22; 8 male, 3 female) and included both Japanese ($n = 5$) and Chinese ($n = 6$) students. All participants had encountered classical Japanese haiku during their compulsory education, and none had specialized training in haiku composition, literary criticism, music theory, or image-generation technologies. Participants viewed the five demonstrations and completed a structured questionnaire after each one, providing ratings on a five-point Likert scale as well as brief free-form comments. It allowed us to collect both quantitative and qualitative feedback on the clarity, harmony, and overall usefulness of the multimodal presentations. This questionnaire did not require ethics committee approval, as it did not involve animal or human clinical trials and did not raise significant ethical concerns. This study adhered to the ethical principles outlined in the Declaration of Helsinki. All participants were informed of the questionnaire's purpose and provided their informed consent before participating. All participants were guaranteed anonymity and confidentiality, and participation was entirely voluntary.

(3) Measures

After viewing each demo, participants evaluated three aspects of the presentation using a five-point Likert scale (1 = strongly disagree, 5 = strongly agree):

- the appropriateness of the generated illustration for the haiku,
- the harmony between the haiku and the background music,
- the overall sense of unity of the harmonious multimodal presentation.

After completing all five demos, participants also provided overall impressions of the system, including perceived novelty, emotional clarity, and general usefulness for haiku appreciation, followed by optional free-form comments. This combination of repeated individual ratings and final overall feedback allowed us to assess both within-demo and across-demo user impressions.

4.5.2. Results

Figure 23 illustrates the distribution of participants' ratings for the five demonstrations (D1 to D5). Across all demos, most scores range between 3 and 5, indicating generally positive impressions of the generated illustrations, the selected background music, and their combined presentation. In particular, D1, D2, D4, and D5 show a concentration of scores around 4. In contrast, D3 shows greater variability in user perception, indicating a wider range of scores.

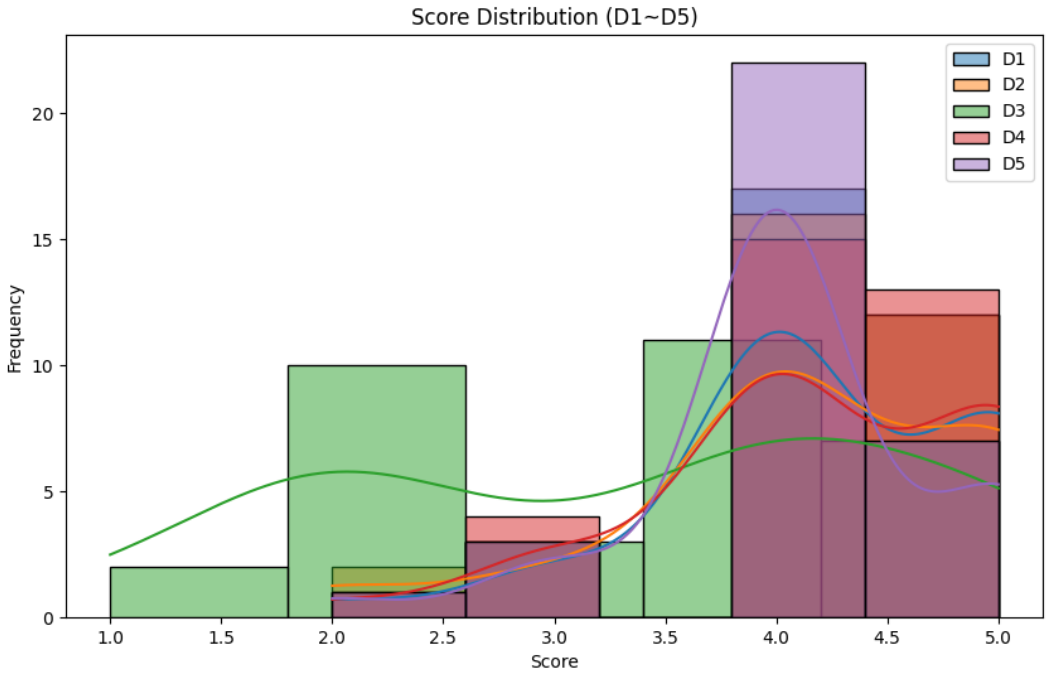

Figure 23. Score distribution for demonstrations (D1 to D5) is shown in the histograms and kernel density estimates. These show the distribution of participant ratings for each demonstration. Scores range from 1 (strongly disagree) to 5 (strongly agree).

Figure 24 summarizes the mean and standard deviation of the five demos. Four demonstrations (D1, D2, D4, D5) achieved mean scores above 4.0, indicating that the majority of participants perceived the multimodal presentations as appropriate and effective. D3 showed the lowest average score (mean = 3.33), consistent with its broader distribution in Figure 23. Standard deviations fall between 0.5 and 0.8, suggesting moderate inter-participant variability.

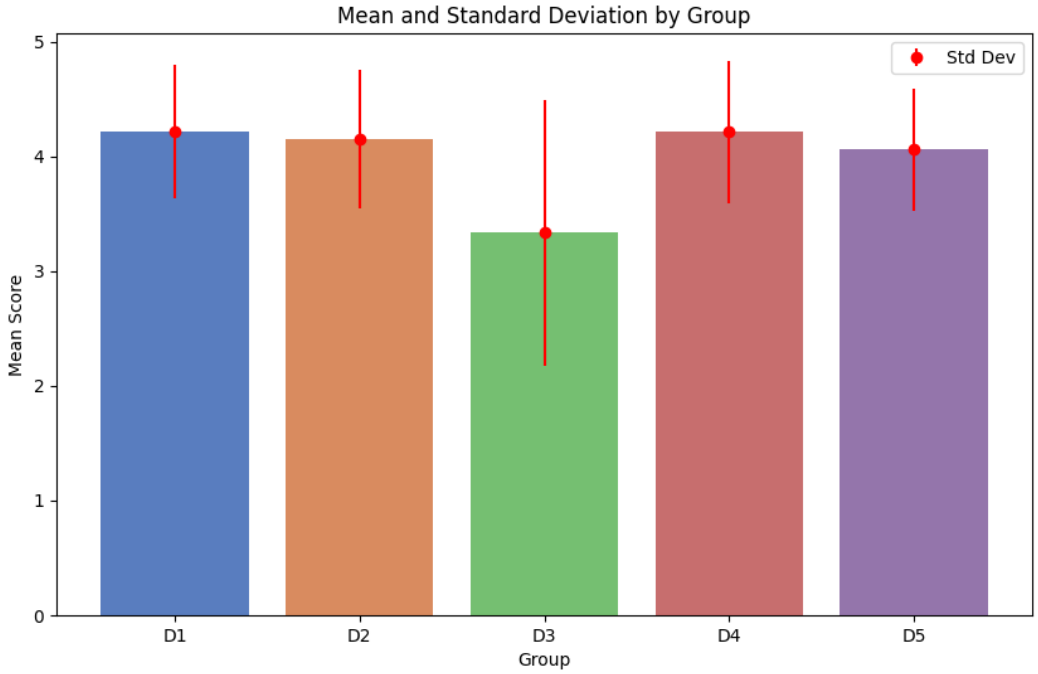

Figure 24. Mean and standard deviation of user ratings for the five demonstrations. Error bars indicate standard deviations.

4.5.3. Summary

The results of Experiment E demonstrate that the proposed multimodal haiku presentation is generally effective and well-received. Participants reported that the combination of text, illustration, and background music helped them intuitively grasp the emotional tone and atmospheric qualities of the haiku. This combination offered a richer appreciation experience than reading the text alone. Four demonstrations (D1, D2, D4, and D5) achieved consistently high scores, further confirming the usefulness of the integrated haiku–image–music presentation.

In contrast, D3 received the lowest mean rating and exhibited the largest variance, indicating weaker alignment between the poem and its accompanying modalities. This pattern is consistent with the broader score distribution shown in Figure 23. Free-form comments also support this interpretation: several participants described the music in D3 as "slightly mismatched," and others felt that the illustration was "less representative of the poetic imagery." These observations suggest that certain poems, particularly those with subtle emotional shifts or less prototypical seasonal cues, may require more flexible or context-sensitive alignment mechanisms.

A likely source of this variability lies in the system's use of fixed weights for *Valence*, *Energy*, and *Dynamism* across all haiku. While this global weighting scheme works well for most poems, it may not fully capture the nuanced affective structures of more complex haiku. User feedback also indicates that culturally grounded cues (e.g., seasonal motifs, genre-specific expression patterns) play an important role in shaping perceived appropriateness.

Taken together, these findings point to promising directions for refinement, including adaptive weighting strategies, improved mapping between textual cues and musical features, and additional constraints in the image-generation pipeline. Experiment E thus complements the analytical validation of Experiments A to D by showing that the proposed cross-modal system is not only structurally valid but also capable of producing multimodal outputs that users find engaging, informative, and helpful for their haiku appreciation.

5. Discussion

The study investigated the potential for representing the aesthetic and emotional dimensions of haiku within a low-dimensional affective space derived from intelligent data analysis. Additionally, it examined whether such representations could facilitate coherent cross-modal alignment between text, music, and imagery. Across a series of analytical and user-centered evaluations (our validation framework), the results suggest that the proposed framework is both structurally sound and perceptually meaningful. At the same time, several limitations reveal significant challenges for modeling poetic content and designing multimodal appreciation systems for haiku appreciation.

5.1. Coherence of the Affective Space

A key finding across the experiments is the emergence of a stable affective structure within the haiku corpus and the music dataset. Despite the brevity and metaphorical nature of haiku, the VED representation derived from our BERT-based intelligent data analysis (Experiment A) produced systematic patterns consistent with the traditional aesthetics of Japanese poetry. Likewise, the VET representation of instrumental music (Experiment C) displayed an interpretable geometry, with *Valence*, *Energy*, and *Tempo* forming meaningful discriminative axes.

The strong cross-modal correspondence observed in the axis-wise correlations (Experiment D-1) indicates that these textual and musical affective spaces are not independent. Instead, they appear to share a compatible emotional geometry. This alignment is noteworthy given that the two modalities were processed separately using different data analysis principles.

5.2. Cross-Modal Alignment and Multimodal Interpretation

The nearest-neighbor matching between haiku and music produced high correlations across all affective dimensions (Experiment D-1) and non-random consistency at the cluster level (Experiment D-2). These results demonstrate that a simple geometric matching scheme, based solely on affective similarity, can generate a harmonious multimodal experience that users perceive as coherent and emotionally appropriate.

The user evaluation (Experiment E) further supports this conclusion. Most demonstrations were rated positively, and free-form comments indicate that participants found the cross-modal outputs helpful for understanding the emotional atmosphere of each poem. These findings suggest that affective similarity, even when derived from minimal feature sets, can serve as an effective mechanism for multimodal haiku appreciation.

5.3. Interpretability and Cultural Variability

Although the overall reception was positive, variability in the ratings (notably for Demo 3 in Experiment E) indicates that affective alignment alone is not always sufficient. Haiku often rely on cultural cues, seasonal conventions, and highly compressed metaphorical associations. General-purpose language models have a challenge in capturing these elements, which can result in mismatches between generated images, selected music, and user expectations.

Cross-cultural interpretation also played a role. The participant group included both Japanese students and international students studying in Japan. Their differing cultural backgrounds occasionally produced divergent impressions of the same multimodal presentation. It suggests that affective alignment may require conditioning based on cultural or experiential factors. It highlights a broader challenge for the cross-cultural application of AI-based art interpretation: emotional resonance is partly universal but also significantly influenced by cultural norms and literary tradition.

5.4. Limitations and Future Challenges

Several limitations of the present study offer opportunities for improvement.

5.4.1. Limited Textual Domain

The haiku corpus used in this study consisted exclusively of the 1067 poems written by Matsuo Bashō. Although Bashō is foundational to the haiku tradition, his works occupy a relatively narrow emotional and stylistic space. To explore the feasibility of supervised emotion classification, we previously conducted a small annotation study in which several participants assigned one of four categorical labels (Joy, Anger, Sadness, Calmness) to each poem. Despite minor individual differences, a striking pattern emerged: the majority of annotators assigned the label "Calmness" to an overwhelming proportion of poems, often exceeding 70%. This implies that a trivial classifier labeling every haiku as "Calmness" would achieve a deceptively high baseline accuracy. It would render supervised categorical emotion prediction uninformative and unsuitable for this genre.

This bias is not unique to Bashō. Haiku as a literary form, and classical Japanese poetry more broadly, tends to gravitate toward subtle, contemplative, and low-arousal affective states. Such tendencies reflect deeper cultural aesthetics such as *mono no aware* (the pathos of transience) and *wabi–sabi* (the beauty of imperfection and austerity). Given these characteristics, categorical affect labels fail to capture the nuanced and predominantly tranquil emotional spectrum of haiku, which motivated our shift toward continuous, low-dimensional affective representations (our VED framework) rather than discrete emotion categories.

However, we acknowledge that relying on a single author introduces a potential stylistic bias. Consequently, the current system may be overfit to Bashō's specific aesthetic preferences. In this study, we intentionally position the system as a controlled proof of concept, focusing on the most representative figure of the genre. Therefore, at this stage, we do not claim generalization across different authors or cultures. Furthermore, this focus on tranquility results in a sparse distribution in the high-arousal regions of the VED space. Future work must incorporate diverse corpora to validate the broader applicability of the proposed representation. Specifically, including *waka* could introduce greater variance in emotional polarity due to its romantic themes. Meanwhile, classical Chinese poetry, which often depicts grandeur or social turbulence, would expand the dynamic range of *Energy*. Validating the VED framework on such diverse corpora is essential to ensure its cross-cultural generalizability beyond the specific aesthetic of Bashō.

5.4.2. Dependence on Non-Curated Music Categories

The music dataset relied on Free BGM DOVA-SYNDROME's "Japanese-style" tag, which includes both traditional-sounding pieces and modern high-energy genres such as rock and metal. Although these styles expand the expressive range, they also introduce noise into the affective feature space. More granular genre control or expert-based music curation may further improve cross-modal matching quality.

5.4.3. Fixed Affective Weighting

The current system uses a fixed weighting scheme for VED/VET similarity. The subjective feedback from Experiment E, especially the variability observed in Demo 3, suggests that users are sensitive to mismatches when poems contain subtle emotional shifts. To address this, future iterations could employ an adaptive weighting mechanism. For example, an attention-based module could dynamically increase the weight of *Dynamism* when high-motion verbs are detected, or prioritize *Valence* when sentiment analysis indi-

cates strong polarity. Such context-aware adjustment would likely reduce misalignment in complex or non-prototypical haiku.

5.4.4. Absence of Expert Baselines

While all participants (Experiment E) were familiar with haiku from school education, none were professional poets or literary specialists. Their interpretations and free-form comments were informative but limited. Future work should compare them with expert judgments to strengthen the validation of literary and aesthetic alignment. Unlike general users who evaluate intuitive harmony, experts can assess the system using specialized metrics, such as "Seasonal Relevance," "Cultural Fidelity" to traditional aesthetics, and "Interpretive Depth." These metrics would help determine whether the system captures the profound subtext of the poem, such as *wabi-sabi*, rather than merely its surface-level semantics.

5.4.5. Prompt-based Image Generation Limits

The generated watercolor images were generally well-received in Experiment E; however, achieving fine-grained control over composition, kigo depiction, and culturally specific motifs remains challenging. Current diffusion-based and instruction-following image models may struggle to capture traditional Japanese aesthetics without additional fine-tuning or style conditioning. Furthermore, as the system relies on large-scale transformer models, the outputs are subject to exposure bias and distributional shift, where the generated content may diverge from the intended distribution of traditional poetic imagery [30]. This highlights the need for more robust generation techniques that can better align with specific cultural domains.

To mitigate these issues, prompt engineering can be refined in several concrete ways. First, prompts may include explicit stylistic descriptors rooted in classical Japanese art, such as *wabi-sabi*, *yūgen*, traditional ink-wash painting, or *ukiyo-e* composition, to provide clearer cultural grounding. Second, the kigo component can be strengthened by linking each seasonal word to its canonical visual motifs (e.g., plum blossoms, harvest moon, first snow), reducing the model's reliance on implicit inference. Integrating a retrieval-augmented generation (RAG) approach or a lookup table grounded in a *Saijiki* (season-word dictionary) would be a practical extension to improve cultural authenticity. Finally, adopting a more structured prompt format with separate slots for emotional tone, kigo motifs, artistic style, and composition could improve reproducibility and aesthetic alignment across generated images.

5.5. Broader Implications and Cross-Cultural Extension Potential

Despite these limitations, the findings demonstrate that affective representations, derived from intelligent data analysis, offer a promising basis for multimodal literary interpretation. The alignment observed in this study suggests that emotional structure can act as a cross-modal bridge between text, sound, and visual imagery, even in highly compact poetic forms such as haiku.

Therefore, this work contributes to the growing body of research that explores the application of intelligent data analysis to literary appreciation, the interpretation of multimodal art, and culturally grounded generative systems. Given this framework's potential for cross-cultural applications, future extensions may incorporate larger haiku corpora, other traditional Japanese poetic forms (such as waka and tanka), and adaptive systems that respond to user preferences and cultural backgrounds. Furthermore, broadening the scope beyond haiku may yield richer affective diversity. Classical Chinese poetry, for example, encompasses a much wider emotional range and a broader variety of themes, offering a valuable testbed for evaluating whether our affective representations generalize across literary traditions.

6. Conclusions

This study investigated whether the aesthetic and emotional qualities of haiku could be expressed through an intelligent data analysis-derived compact affective space and used to create a harmonious, multimodal experience involving text, music, and imagery. By integrating BERT-based emotional features (our intelligent data analysis component), acoustic affect descriptors, and controlled prompt-driven illustration, the proposed system demonstrates that affective similarity alone can serve as a viable application for enhancing cross-modal poetic presentation.

Through five complementary experiments, we validated the system from multiple perspectives. Experiment A demonstrated that the VED representation extracted from BERT-based contextual embeddings captures interpretable emotional structure within the Bashō haiku corpus. Experiment B confirmed that BERT provides superior stability and semantic fidelity compared with TF-IDF and Word2Vec as the basis for this analysis. Experiment C established the reliability of the musical VET descriptors, and Experiment D revealed strong cross-modal structural consistency between the textual and musical affective spaces. Finally, Experiment E demonstrated that the combined multimodal outputs were positively evaluated by users, indicating that affect-guided alignment enhances intuitive poetic appreciation.

These findings collectively suggest that low-dimensional affective spaces, when derived from intelligent data analysis, provide a practical, interpretable, and modality-independent foundation for applications in multimodal poetic appreciation. Although the current validation used Bashō's haiku as a controlled proof of concept, the proposed methodology is adaptable. It operates as a general, modality-independent framework that can coherently map text, sound, and visual imagery.

Several promising directions emerge from this work. To address the limitation of the single-author dataset and realize the full cross-cultural extension potential of this framework, future extensions may incorporate larger and more diverse poetic corpora, including waka, tanka, and classical Chinese poetry, to broaden the affective range and reduce genre-specific bias. More adaptive alignment mechanisms, capable of adjusting feature weights based on poem-specific cues or user preferences, may further improve cross-modal coherence. Advances in vision-language models also offer opportunities for more faithful and culturally sensitive illustration. Finally, expanding the validation to include expert poets, literary scholars, and diverse cross-cultural audiences will help establish a more comprehensive understanding of multimodal poetic appreciation.

Overall, this work provides methodological insights and empirical evidence indicating that affect-guided multimodal generation could be a valuable approach for intelligent data analysis applications in poetic appreciation. The developed framework is an initial step toward systems integrating emotion, culture, and creativity across modalities to support more accessible and engaging poetic art experiences.

Author Contributions: Conceptualization, R.F. and Y.W.; methodology, R.F.; software, R.F.; validation, R.F. and Y.W.; formal analysis, R.F.; investigation, R.F.; resources, R.F.; data curation, R.F.; writing—original draft preparation, R.F.; writing—review and editing, R.F. and Y.W.; visualization, R.F.; supervision, Y.W.; project administration, Y.W.; funding acquisition, Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by JSPS KAKENHI Grant Numbers JP21K17862 and JP25K15354.

Institutional Review Board Statement: Ethical review and approval were waived for this study due to its nature as an anonymous user questionnaire survey, which does not involve personal data collection or invasive procedures, in accordance with the "Ethical Guidelines for Medical and Biological Research Involving Human Subjects" established by the Japanese Ministry of Health, Labour, and Welfare (MHLW) and the "Regulations for General Research Involving Human Subjects" at Yamaguchi University.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The datasets used in this study are openly available as follows: The bashoDB (Bashō Haiku Database) provided by Yamanashi Prefectural University, available at http://www2.yamanashi-ken.ac.jp/~itoyo/basho/basho.htm (accessed on 18 November 2025); and the Free BGM DOVA-SYNDROME music library, available at https://dova-s.jp/ (accessed on 18 November 2025). Further inquiries can be directed to the corresponding author.

Acknowledgments: The authors gratefully acknowledge the providers of the publicly available datasets and tools used in this study. These include the bashoDB (Bashō Haiku Database) from the University of Yamanashi, seasonal term dictionaries from Weblio, BERT models from Tohoku University, and music from Free BGM DOVA-SYNDROME.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
ARI	Adjusted Rand Index
BPM	Beats-per-minute
LLMs	Large language models
MER	Music emotion recognition
MFCCs	Mel-frequency cepstral coefficients
NLP	Natural language processing
PCA	Principal component analysis
RAG	Retrieval-augmented generation
RMS	Root-mean-square
STFT	Short-time fourier transform
ZCR	Zero-crossing rate

Appendix A. Haiku Used in the Image-Reproducibility Experiment

Table A1 lists the six representative haiku used in Experiment D-3 (Section 4.4.3). These haiku were chosen to cover all six seasonal categories.

Table A1. Haiku metadata used for constructing prompts in Experiment D-3.

ID	Original Text	English Translation	Season Name	Season-Specific Scene	Season-Specific Color	Cluster-Specific Emotional Tone
839	古池や蛙飛びこむ水の音	Old pond—a frog leaps in, the sound of water.	Spring	A tranquil spring landscape with cherry blossoms and fresh greenery.	Soft pastel pink and light green.	High-energy, tense mood.
648	夏草や兵どもが夢の跡	Summer grasses—all that remains of warriors' dreams.	Summer	A bright summer scene near water, with deep blue sky and strong sunlight.	Vivid blue and warm yellow.	Low-energy, calm mood.

Table A1. *Cont.*

ID	Original Text	English Translation	Season Name	Season-Specific Scene	Season-Specific Color	Cluster-Specific Emotional Tone
932	名月や池をめぐりて夜もすがら	Bright autumn moon—circling the pond all through the night.	Autumn	An autumn landscape with maple leaves and a slightly nostalgic atmosphere.	Deep orange, red and brown.	Low-energy, calm mood.
78	いざさらば雪見にころぶ所まで	Come, let us go—until we tumble in the snow while viewing it.	Winter	A quiet winter scene with snow and a cold, clear atmosphere.	White, pale blue and muted gray.	High-energy, tense mood.
231	徒歩ならば杖突坂を落馬かな	Had I been on foot, I would not have fallen from my horse on Tsue-tsuki Slope.	All Year	A simple natural landscape that could belong to any season.	Balanced natural colors.	Low-energy, calm mood.
234	門松やおもへば一夜三十年	New Year's pine—thinking back, one night feels like thirty years.	New Year	A calm New Year scene with subtle festive atmosphere.	Soft red and white.	High-energy, tense mood.

References

1. Rowland, P. New Directions in English-language Haiku: An Overview and Assessment. *Iafor J. Lit. Librariansh.* **2013**, *2*, 53–66. [CrossRef]
2. Kawamura, H.; Yamashita, M.; Yokoyama, S. *Artificial Intelligence Reads Haiku: The Challenge of AI Issa-kun*; Ohmsha, Ltd.: Tokyo, Japan, 2021.
3. Yokoyama, S.; Yamashita, T.; Kawamura, H. Generation and Selection of Haiku Poems Using Deep Learning. *J. Jpn. Soc. Artif. Intell.* **2019**, *34*, 467–474. [CrossRef]
4. Yokoyama, S.; Yamashita, T.; Kawamura, H. Development of a Selection Mechanism Using Deep Auto-Regression Model with Examples of Seasonal Fixed-form Haiku in Japanese. In Proceedings of the 38th Annual Conference of the Japanese Society for Artificial Intelligence, Online, 28–31 May 2024. [CrossRef]
5. Yuki, T.; Ueda, N.; Oka, T.; Komachi, M. Toward Automatic Evaluation of Haiku Using GPT. In Proceedings of the 38th Annual Conference of the Japanese Society for Artificial Intelligence, Online, 28–31 May 2024. [CrossRef]
6. Hanano, A.; Yokoyama, S.; Yamashita, T.; Kawamura, H. Evaluation of Haiku using Masked Language Model RoBERTa. In Proceedings of the 84th National Convention of the Information Processing Society of Japan, Online, 3–5 March 2022; Volume 2022; pp. 1061–1062.
7. Tateishi, K.; Mizuguchi, S. Are we biased against AI-made haiku poems? *Proc. Linguist. Soc. Am.* **2025**, *10*, 1–15. [CrossRef]
8. Russell, J.A. A Circumplex Model of Affect. *J. Personal. Soc. Psychol.* **1980**, *39*, 1161–1178. [CrossRef]
9. Shahriar, S.; Al Roken, N.; Zualkernan, I. Classification of Arabic Poetry Emotions Using Deep Learning. *Computers* **2023**, *12*, 89. [CrossRef]
10. Walunj, V.; Choudhary, A.; Kale, A.; Gade, M. Emotion Recognition from Formal Text (Poetry). *Int. J. Adv. Res. Comput. Commun. Eng.* **2023**, *12*, 1469–1473. [CrossRef]
11. Ohmameuda, T. Preliminary Study on Automatic Extraction of Seasonal Words (Kigo) in Haiku. *Gunma-Kosen Rev.* **2023**, *42*, 95–100. [CrossRef]
12. Kim, Y.; Schmidt, E.; Migneco, R.; Morton, B.; Richardson, P.; Scott, J.; Speck, J.; Turnbull, D. Music emotion recognition: A state of the art review. In Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR, Utrecht, The Netherlands, 9–13 August 2010; pp. 255–266.
13. Yang, Y.H.; Chen, H.H. *Music Emotion Recognition*; CRC Press: Boca Raton, FL, USA, 2012. [CrossRef]
14. Zhao, J.; Yoshii, K. Multimodal Multifaceted Music Emotion Recognition Based on Self-Attentive Fusion of Psychology-Inspired Symbolic and Acoustic Features. In Proceedings of the 2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Taipei, Taiwan, 31 October–3 November 2023; pp. 1641–1645. [CrossRef]
15. Cui, X.; Wu, Y.; Wu, J.; You, Z.; Xiahou, J.; Ouyang, M. A review: Music-emotion recognition and analysis based on EEG signals. *Front. Neuroinformatics* **2022**, *16*, 997282. [CrossRef] [PubMed]
16. Zhao, Y.; Yang, M.; Lin, Y.; Zhang, X.; Shi, F.; Wang, Z.; Ding, J.; Ning, H. AI-Enabled Text-to-Music Generation: A Comprehensive Review of Methods, Frameworks, and Future Directions. *Electronics* **2025**, *14*, 1197. [CrossRef]
17. Mao, Z.; Zhao, M.; Wu, Q.; Zhong, Z.; Liao, W.H.; Wakaki, H.; Mitsufuji, Y. Cross-Modal Learning for Music-to-Music-Video Description Generation. *arXiv* **2025**, arXiv:2503.11190.

18. Minato, K.; Onai, R. A prototype system of composite image generation for Haiku. *Ite Tech. Rep.* **2009**, *33*, 43–46. (In Japanese) [CrossRef]
19. Minato, K.; Okabe, M.; Onai, R. A Prototype System for Generating Composite Images Suitable for Haiku. In Proceedings of the Workshop on Interactive Systems and Software (WISS), Hong Kong, China, 12–14 December 2010; Information Processing Society of Japan (IPSJ): Tokyo, Japan, 2010; p. D37.
20. Toda, K.; Ito, A.; Yoshino, T. Investigation of an Automatic Generation System of Images Adapted to Haiku Scenes. In Proceedings of the 2023 IPSJ Kansai Chapter Annual Convention, Osaka, Japan, 19 August 2023; Volume 2023, 7p.
21. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-Shot Text-to-Image Generation. *arXiv* **2021**, arXiv:2102.12092.
22. Sasaki, S.; Ushiama, T. A Method of Generating Images to Represent Impressions of a Piece of Music Using the Text2Image Model. In Proceedings of the IEICE Technical Committee on Data Engineering, Kitakyushu International Conference Center, Sapporo, Japan, 29 November–1 December 2023; IEICE Technical Report: DE2023-26 (No. DE-192); Volume 123, pp. 89–94.
23. Liu, B.; Fu, J.; Kato, M.P.; Yoshikawa, M. Beyond Narrative Description: Generating Poetry from Images by Multi-Adversarial Training. In Proceedings of the ACM International Conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; ACM: New York, NY, USA, 2018; pp. 783–791. [CrossRef]
24. Tosa, N. *Cross-Cultural Computing: An Artist's Journey*; Springer Series on Cultural Computing; Springer: London, UK, 2016. [CrossRef]
25. Ohno, M.; Yokosawa, K.; Narumi, T. Basics of Integrated Perception and its Application to the Multi-Sensory Interface. *Trans. Virtual Real. Soc. Jpn.* **2022**, *27*, 18–28. [CrossRef]
26. Lu, M.Y.; Chen, B.; Williamson, D.F.K.; Chen, R.J.; Zhao, M.; Chow, A.K.; Ikemura, K.; Kim, A.; Pouli, D.; Patel, A.; et al. A Multimodal Generative AI Copilot for Human Pathology. *Nature* **2024**, *634*, 466–473. [CrossRef] [PubMed]
27. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; Inui, K., Jiang, J., Ng, V., Wan, X., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 3982–3992. [CrossRef]
28. Ethayarajh, K. How Contextual Are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; Inui, K., Jiang, J., Ng, V., Wan, X., Eds.; pp. 55–65. [CrossRef]
29. Eerola, T.; Vuoskoski, J.K. A comparison of the discrete and dimensional models of emotion in music. *Psychol. Music* **2011**, *39*, 18–49. [CrossRef]
30. Pozzi, A.; Incremona, A.; Tessera, D.; Toti, D. Mitigating exposure bias in large language model distillation: An imitation learning approach. *Neural Comput. Appl.* **2025**, *37*, 12013–12029. [CrossRef]