2025 IEEE World Haptics Conference (WHC)
Suwon Convention Center, Suwon, Korea
July 8 ~ 11, 2025

Text-Driven Generative Framework for Multimodal Visual and Haptic Texture Synthesis

Myrah Naeem*
Dept. of Computer Science and Engineering, Kyung Hee University
Yongin, South Korea
ORCID: 0000-0002-1516-4660

Mudassir Ibrahim Awan*
Dept. of Computer Science and Engineering, Kyung Hee University
Yongin, South Korea
ORCID: 0000-0002-1825-0097

Seokhee Jeon†
Dept. of Computer Science and Engineering, Kyung Hee University
Yongin, South Korea
ORCID: 0000-0002-0413-9646

***Abstract*—This paper presents a novel framework for generating both visual and haptic textures from user-provided text descriptions. The proposed text-to-haptic pipeline combines generative AI with data-driven tactile rendering to enable intuitive and perceptually accurate texture synthesis. A text-to-image model (i.e., Stable Diffusion) generates high-quality visual representations of textures from descriptive text prompts. These visual textures are processed through a regression-based deep learning architecture, termed AttributeNet, which predicts perceptual attributes, such as roughness and softness, mapping them onto continuous perceptual scales. Finally, an interpolation-based texture authoring algorithm synthesizes vibrotactile signals based on the predicted attributes, enabling us to render haptic feedback aligned with the visual and textual input. To the best of our knowledge, this is the first complete framework to generate visual and haptic texture signals based on text-based inputs. AttributeNet's haptic attribute predictions achieved improved accuracy over existing methods, and a user study further validated the framework, with participants favoring its quality and usability.**

***Index Terms*—Haptic Texture, Text-to-Haptics, Tactile Feedback, Deep Learning, Generative AI, Multimodal Interaction.**

I. Introduction

Among various haptic attributes, haptic texture is one of the most critical for humans to perceive both the functional and aesthetic qualities of surfaces [1]–[3]. It conveys sensations of roughness, hardness, and slipperiness by generating multiple signals related to contact dynamics between the finger or tool and small-scale surface features [4], [5]. These signals typically consist of rapidly fluctuating pressure (either at a single point or across a 2D area, depending on contact type) during sliding interactions or a quasi-static pressure distribution map in static contact. Ideally, generating such signals requires both an accurate, high-speed simulation of contact dynamics and a high-resolution surface geometry model. However, this approach is not optimal for real-time rendering [6].

Recent research in haptic texture modeling/rendering has primarily relied on data-driven approaches: related haptic signals are recorded, modeled, and interpolated to reproduce them. In order to avoid complexity, most of the work assumed tool-mediated interaction, which only needs the recording of vibration [7]. These vibrations vary and are unique to each surface when different interaction parameters are applied (e.g., scanning speed and applied force) [8]. A significant number of researchers have modeled these textures by recording vibrations as acceleration data, rendering them through tool-based displays where surface-specific vibrations are synthesized based on interactions [9], [10]. These modeling techniques include stochastic modeling [11], [12] and deep learning-based synthesis [3], [13], [14]. Furthermore, state-of-the-art texture libraries derived from a large number of diverse textures using these approaches have been built [9], [15], making the modeling and rendering of haptic texture readily available for researchers in various applications [11], [16]. However, one of the main drawbacks of data-driven methods is that they can only generate what has been explicitly modeled.

Consequently, researchers have recently turned their attention towards generative methods capable of creating new textures that are not modeled, aligning with user-defined perceptual attributes such as softness, roughness, and bumpiness. Most existing methods rely heavily on perturbing or interpolating textures from libraries [17] and have achieved significant accuracy in producing perceptually correct textures. Recently, a few recent studies have adopted deep learning-based approaches [18], such as generative adversarial networks (GANs). While promising, these approaches have limitations, including the need for large training datasets and high computational costs, making real-time rendering and texture generation challenging [19]. Consequently, interpolation-based texture authoring remains the preferred state-of-the-art approach for new texture synthesis [17].

Despite the advancements in texture generation, both interpolation-based and deep learning-based approaches have largely overlooked two pivotal aspects. First, they do not integrate visual representations of textures, which could greatly enhance user engagement and usability [17], [18]. Second, they lack mechanisms for interfacing with users, i.e., incorporating user inputs such as textual or verbal descriptions to generate textures with desired perceptual and visual attributes [17]–[20]. For instance, a user may wish to create a texture described as "a rough, soft carpet-like surface" with specific

This research was supported by the National Research Council of Science & Technology (NST) grant by the Korea government (MSIT) (CRC23021-000) and by the MSIT(Ministry of Science and ICT) Korea under the Mid-Researcher Program (2022R1A2C1008483) supervised by the NRF Korea.

*These authors contributed equally to this work. †Seokhee Jeon is the corresponding author (e-mail:jeon@khu.ac.kr).

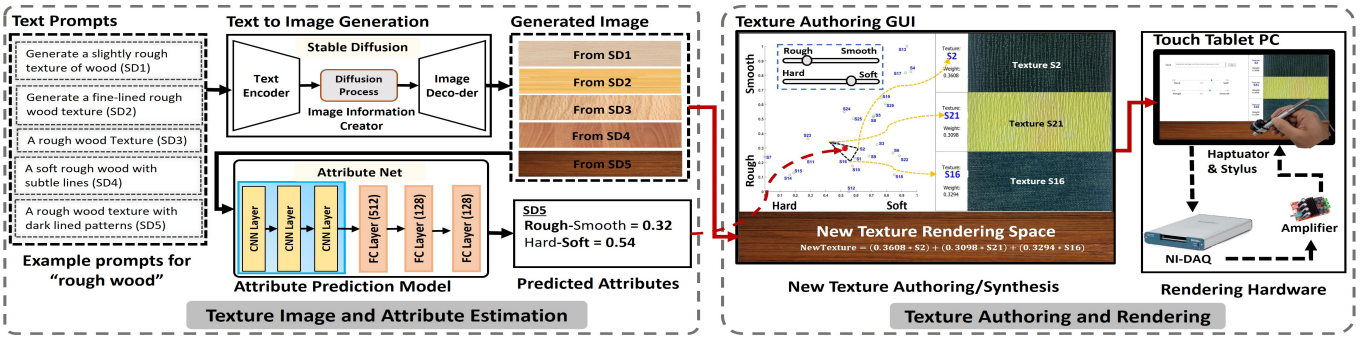


Fig. 1. The proposed framework for generating visual and haptic textures from text prompts. User-provided text prompts are processed by a Stable Diffusion model to generate a corresponding image, which is then passed to a haptic attribute prediction model (AttributeNet) to estimate hard-soft and rough-smooth levels. These attributes are mapped onto a 2D-Texture Authoring Space [17] to synthesize required texture via interpolation of vibration signals from a pre-built texture library [15]. Users interact with the system on a touch surface, experiencing both visual and haptic feedback through a stylus equipped with a haptuator. Additionally, sliders are provided for each attribute to refine and enhance the haptic texture feedback further.

visual and tactile properties as indices. This gap highlights the need for algorithms capable of translating descriptive inputs into textures with corresponding visual and haptic feedback.

Motivated by these challenges, we propose a novel framework that bridges generative modeling and haptic rendering for texture synthesis. With advancements in generative artificial intelligence, particularly text-to-image synthesis models like Stable Diffusion [21], new possibilities emerge to unify visual and haptic texture generation. This study introduces a text-to-haptic framework that leverages diffusion models to generate visual textures from descriptive inputs and couples them with data-driven tactile rendering [17] for haptic feedback.

The proposed pipeline consists of three key stages. Initially, a text-to-image model such as Stable Diffusion [21] generates high-quality visual representations of textures from user-defined text prompts. Subsequently, a deep learning-based architecture, termed AttributeNet, employing ResNet [22] as its backbone, predicts the perceptual attributes of textures, such as roughness and softness, from the generated images. Finally, an interpolation-based texture authoring algorithm [17], integrated with a texture library [15] and rendering techniques [23], generates tactile vibrations based on the predicted perceptual attribute values. This integration of generative AI with haptic texture authoring enables the framework to synthesize perceptually accurate haptic textures from textual descriptions. To the best of our knowledge, this is the first implementation of an end-to-end, text-driven haptic texture generation framework capable of synthesizing both visual and tactile outputs. This text-based approach can also be enhanced by integrating voice-to-text models, further expanding its potential applications. By bridging generative visual models with haptic feedback, this study introduces a novel direction for texture generation, advancing the capabilities of existing texture libraries and enabling greater flexibility for VR, AR, and HCI applications.

II. Proposed Framework

The proposed framework, illustrated in Fig. 1, generates both visual and haptic textures from user-provided textual descriptions. It operates as a sequential pipeline comprising two main components. The first component utilizes a text-to-image generative model (i.e., Stable Diffusion) to create visual textures and predicts corresponding haptic attribute values (hard-soft, rough-smooth) from the generated images. The second component synthesizes haptic textures by mapping the predicted attributes to vibration signals that replicate the tactile properties of the desired texture. This approach delivers a multi-modal experience, combining visual and haptic feedback from textual input. The following subsections detail the key stages of the framework, including text-to-image generation, the texture authoring algorithm, haptic attribute prediction, and the overall rendering process and hardware setup.

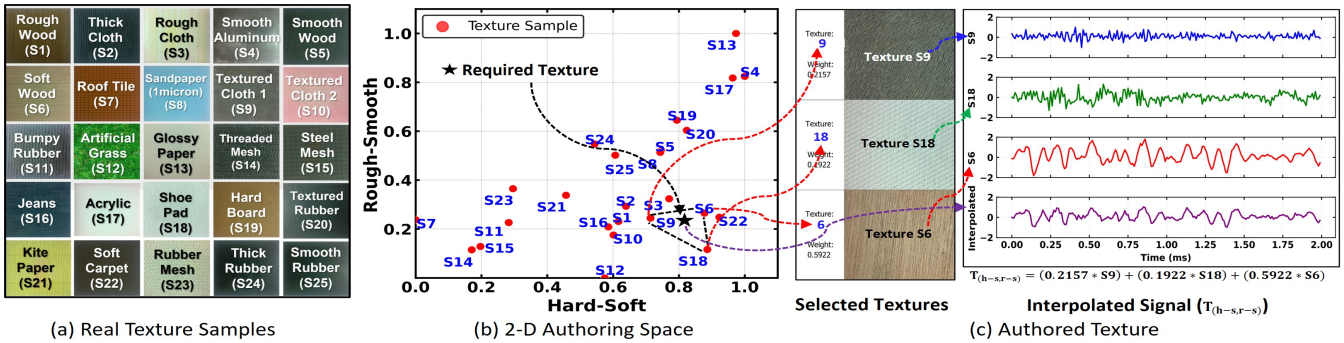
A. Text to Texture Image Generation

Recent advancements in generative AI have led to growing interest in text-to-image models due to their intuitive and expressive capabilities for generating desired outputs [24]–[26]. Among several text-to-image open-source models, including GLIDE [27] and DALL-E [28], Stable Diffusion [21] has emerged as a powerful tool, particularly for tasks like photorealistic image generation and texture synthesis for VR/AR applications [25]. Trained on the extensive LAION-5B dataset [29], which includes real and synthetic textures, clothing, and various objects, Stable Diffusion (SD) offers high adaptability for diverse use cases.

For our framework, we therefore selected SD (v2.1) as the backbone of the text-to-image generation module to synthesize visual textures. Using the diffusers library [30], the model was set to generate 512×512 pixel images to ensure high-quality outputs. It is well established that prompt formulation plays a pivotal role in text-to-image generation, as the specificity and clarity of input text greatly affect the quality and relevance of generated images [24]. To ensure the generated images represent structured textures rather than unrelated visuals (e.g., landscapes), we used a concatenation approach, augmenting prompts with fixed descriptors (i.e., real plain texture, zoomed-in, top view, realistic, no curl, and high definition). These descriptors, refined through experimentation, should henceforth always be included in every text prompt to ensure the generated textures align with task requirements. This approach remained consistent and was used to generate synthetic images for training AttributeNet (Sec.II-C) and during perceptual experiments at the backend (Sec.III-B).

B. Haptic Texture Authoring

Haptic texture authoring refers to the process of creating virtual textures with desired tactile perceptions. Two primary



(a) Real Texture Samples (b) 2-D Authoring Space (c) Authored Texture

Fig. 2. (a) 25 real texture samples used to create authoring space. (b) 2D authoring space mapping textures along the Hard-Soft and Rough-Smooth dimensions. (c) Illustration of the synthesized tactile signal using the authoring space, generated by interpolating neighboring textures based on perceptual attributes.

data-driven techniques are commonly employed for this purpose: deep learning-based approaches [18], and interpolation-based techniques [17]. While deep learning methods, such as generative adversarial networks (GANs), have been explored for texture synthesis, they are computationally expensive, data-intensive, and remain underexplored in the haptic domain [19]. In contrast, interpolation-based techniques, as proposed by Hassan et al. [17], offer a computationally efficient alternative and achieve high perceptual accuracy of 94% for authored texture. It can generate new textures by interpolating between existing texture models adopted from texture library [15] based on input perceptual attribute values, specifically rough-smooth (R-S) and hard-soft (H-S). Given the need for immersive and real-time generation of haptic textures in this study, we adopted the interpolation-based technique proposed by [17].

This interpolation-based approach relies on constructing a structured authoring space (see Fig. 2), integrating two key components: the affective space and the haptic model space. The affective space is developed through psychophysical experiments that explore human perception of texture. Participants evaluate real textures using bipolar adjective scales (e.g., rough-smooth, hard-soft), allowing researchers to map textures into a two-dimensional space where each axis corresponds to a specific perceptual dimension. Multidimensional scaling and regression are employed to establish affective dimensions, ensuring that the textures are positioned meaningfully within this perceptual framework. The haptic model space is constructed using tool-mediated acceleration signals recorded during controlled interactions with surfaces. Parameters such as sliding velocity and normal force are varied systematically to capture the physical properties of textures. Features derived from these acceleration patterns, including Mel Frequency Cepstral Coefficients (MFCC), are extracted and reduced using statistical techniques such as sequential forward selection and principal component analysis (PCA) to isolate dimensions that correlate strongly with the affective axes.

These two spaces are combined to create a unified 2D authoring space that correlates perceptual attributes with physical signal features. This space allows textures to be positioned based on both their perceptual and physical characteristics. To synthesize a new texture, the algorithm maps the desired perceptual values into the authoring space, identifies the three nearest textures using Delaunay triangulation, and calculates interpolation weights based on the Euclidean distances to these points. The final texture signal is computed as a weighted sum:

$$\mathrm{T}_{(h-s,r-s)} = (w_1 \times S_a) + (w_2 \times S_b) + (w_3 \times S_c) \quad (1)$$

where, $\mathrm{T}_{(h-s,r-s)}$ is the synthesized acceleration/vibration pattern for the input perceptual attributes $(h-s, r-s)$, and w_1, w_2, and w_3 are weights determined by the distances to the three nearest points/textures S_a, S_b, and S_c.

For this study, the authoring space was built using 25 real texture samples, including natural and artificial materials like wood and rubber, covering diverse tactile sensations (Figure 2(a)). The constructed 2D space, shown in Figure 2(b), organizes textures along the Rough-Smooth (R-S) and Hard-Soft (H-S) dimensions. An example of texture synthesis using the authoring space is illustrated for $\mathrm{T}_{(0.82,0.23)}$ in Figure 2(c). Initial three plots show the actual acceleration patterns for the real texture models, while the final plot presents the weighted synthesized acceleration for the authored virtual texture. Notably, the texture dataset including the authoring space with perceptual attributes is adopted from [17], and is used throughout the study, including for training AttributeNet (Sec. II-C).

C. Haptic Attribute Estimation

This study enables tactile experiences from user-defined prompts, bridging visual and haptic modalities. While text-to-image models (Sec. II-A) generate visual textures, and tactile feedback is generated by the authoring algorithm (Sec. II-B), a critical step involves predicting haptic attributes from images, linking both modalities. There are two potential approaches to estimating these attributes: either from textual descriptions or generated images. Text-based prediction, despite relying on language models and large labeled datasets, may struggle to capture fine-grained haptic semantics and user intent, leading to ambiguity and inconsistency that could limit accuracy [31].

Conversely, image-based prediction offers a structured representation, as visual texture features inherently correlate with haptic attributes such as roughness, hardness, and bumpiness [32]. Thus, this study adopts an image-based approach employing a CNN-based architecture for haptic attribute prediction. Unlike existing methods that predict user ratings for specific attributes [5], [33], the proposed model maps the physical signal space to the authoring space, ensuring perceptual relevance and compatibility with tactile authoring systems. This approach aligns with the study's goal of generating perceptually accurate textures that align with user intent.

1) Physical Texture Space: The physical texture space was constructed using a combination of real and synthetic datasets. Twenty-five distinct real-world textures were selected (see Fig. 2 (a)). Images were captured using a DP2 Quattro SIGMA digital camera, mounted on a tripod at a fixed height of 300 mm, and resized to a final resolution of 512×512 pixels for uniform processing. These real textures were the same as those used to create the authoring space (Sec. II-B). To enhance the dataset, synthetic images were generated using Stable Diffusion with prompts aligned to real texture properties (see Sec. II-A). Fig. 1 illustrates examples of text prompts and their corresponding generated images. For instance, the prompt "a rough wood texture", along with descriptors, was used to generate variations of a real "rough wood (S1)" sample. Five synthetic images were generated per real texture, resulting in a total of 150 images (25 real + 125 synthetic). This augmentation introduced controlled variability, improving the generalization capability of the haptic attribute prediction model, particularly for handling generative model outputs.

2) 2D Authoring Space: The 2D authoring Space is the same as illustrated in Fig. 2(b), where the 25 textures are placed based on 2D space while preserving their perceptual and tactile signal information in a way so that they are placed on a continuous space that can synthesize required perceptually correct new textures [17]. For the haptic attribute prediction model, each texture was assigned two labels: one for hard-soft (h-s) and one for rough-smooth (r-s), corresponding to its position in the authoring space. Synthetic images generated using Stable Diffusion were assigned the same attribute values as their corresponding real textures. To simulate perceptual variability and improve model robustness, a small random noise (± 0.01) was added to the attribute values. This approach helps prevent overfitting, and enhances generalization, ensuring that the model can learn robust mappings between images and haptic attribute values.

3) AttributeNet: The perceptual attribute prediction model was designed to estimate texture attributes from raw texture images. A CNN-based architecture was selected due to its proven effectiveness in various texture classification [34]–[36] and haptic texture attribute prediction [5], [33].

The proposed architecture is inspired by ResNet-50, a well-known deep CNN model introduced in [22] and pre-trained on the ImageNet dataset, which contains over 1.2 million labeled images across 1,000 classes. To adapt ResNet-50 for haptic attribute prediction, the final fully connected (FC) layer was replaced with a custom sequence of three new FC layers, denoted as FC1, FC2, and FC3, followed by a regression output layer. The sizes of these layers were set to 512, 128, 128, and 2 units, respectively, where the two outputs correspond to the predicted haptic attributes. The convolutional layers from ResNet-50, pre-trained on ImageNet, were retained as feature extractors, while the newly added FC layers were initialized with random weights. Leveraging pre-trained networks is a widely used strategy in the haptics domain as it enhances generalizability and improves performance, particularly when training datasets are limited [35], [37].

The model input consisted of preprocessed RGB texture images resized to 224×224 pixels to match the input requirements of ResNet-50.ReLU was used as the activation function in all intermediate layers, consistent with the original ResNet-50 design. A linear activation function was applied to the final output layer to enable continuous attribute prediction. The model was implemented and trained using the Keras-TensorFlow framework. The Adam optimizer (learning rate = 0.001) was employed to minimize the Root Mean Squared Error (RMSE) loss function. Training ran for 100 epochs, with early stopping (patience = 10 epochs) to prevent overfitting.

D. Haptic Texture Rendering

The proposed framework (Fig. 1) systematically integrates visual and haptic components for texture rendering. The process begins with a user-provided text input, which is passed to the text-to-image model. The generated image is then processed by the haptic attribute prediction model (AttributeNet) to estimate Hard-Soft and Rough-Smooth $(h-s, r-s)$ values. These values serve as input to the texture authoring algorithm [17], which utilizes a haptic texture library [20] and a texture rendering algorithm [23]. The algorithm selects three neighboring textures based on their attributes, and their respective haptic models, accounting for user-applied speed and force [23] upon interaction with the virtual texture, are interpolated to synthesize the texture as illustrated in Fig. 2.

To provide a comprehensive, end-to-end framework, a graphical user interface (GUI) was developed to seamlessly facilitate user interaction, as shown in Fig. 3. It features a text input field (top-left), three selected neighboring texture images (top-right) based on $(h-s, r-s)$, and the generated texture image (bottom). Users can experience real-time haptic feedback from all textures upon interaction. To further enhance user engagement and control, two adjustable sliders for R-S and H-S enable dynamic customization of the generated signal. These sliders modify the perceptual attribute point $(h-s, r-s)$ within the authoring space, which remains hidden from users. Each slider spans the full attribute range, with extreme values representing maximum roughness (0 for R-S) and maximum hardness (0 for H-S). This feature was particularly targeted at improving user immersion and feedback and overcoming any discrepancy in attribute estimations by AttributeNet as well as for user interactivity, as confirmed in Sec. III).

The system is deployed on a tablet PC with an active stylus (Surface Pro 4 and Surface Pen, Microsoft), enabling user interaction, text input, slider adjustments, and capturing speed and force during virtual texture interaction. Haptic feedback is provided by a voice-coil actuator (Haptuator MM1C; Tactile Labs) mounted on the stylus, driven by signals from the texture authoring algorithm (Eq. 1). These signals are dynamically compensated, following [12], to account for the actuator's frequency response before transmission to an NI-DAQ (USB-6351; National Instruments). An amplifier regulates the signal strength between the DAQ and the haptuator. The overall setup is illustrated in Fig. 3 and aligns with rendering methodologies established in prior literature [20].

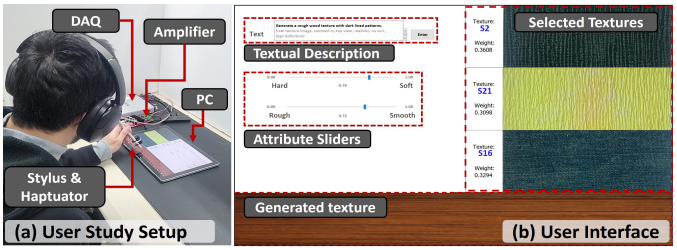


Fig. 3. User study setup (left) and the interface presented to participants during the psychophysical experiment (right).

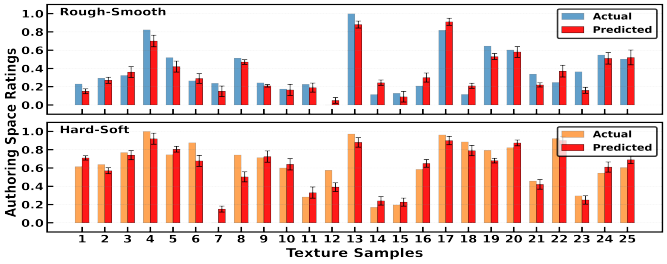


Fig. 4. Comparison of actual vs. predicted values from AttributeNet.

III. Evaluation

To evaluate our first-of-its-kind text-to-haptic texture generation framework, two primary evaluations were conducted: the prediction accuracy of haptic attributes from generated texture images, and a cross-modal user study assessing the perceptual consistency between visual textures and haptic feedback.

A. Prediction of Haptic Attributes

To assess AttributeNet's robustness and generalization, we employed the leave-one-out cross-validation technique, as recommended in similar studies [5], [33]. In this approach, all texture dataset instances were used for training, leaving one out for testing. The model was trained on 24 real texture images and their corresponding synthesized images from Stable Diffusion (SD), yielding 144 training samples. Evaluation was performed on six images: one real and five SD generated. Additionally, we tested the model without SD-images, using only 24 images to assess performance with limited real data and to analyze the effect of SD-images. Each scenario involved 25 training repetitions, totaling 50 runs across both cases.

For further validation, we compared AttributeNet with VisionNet [35], 1D-CNN [33], and VisualCNN [34]. All models were trained under conditions as proposed by respective authors, with the final layer adjusted to predict the two haptic attributes. The Mean Absolute Error (MAE) was selected as the primary metric, as it directly reflects prediction accuracy and provides a clear assessment of model effectiveness [33].

1) Results and Analysis: Fig. 4 compares actual and predicted values for Rough-Smooth (R-S) and Hard-Soft (H-S), where each predicted value represents the mean from six images, incorporating both real and SD-generated textures. In most cases, the actual and predicted values align well, demonstrating the effectiveness of the approach. Table I presents MAE as a percentage, measured on a scale of 0 to 100, across all test samples. AttributeNet outperforms other models, achieving an MAE of 9.87% for R-S and 9.21% for H-S when trained solely on real images. The inclusion of SD-generated images further improves R-S to 7.47% and H-S to 8.16%, while VisionNet and 1D-CNN show higher errors. Empirically, H-S exhibits larger errors than R-S, suggesting that image-based models struggle with compliance-related attributes, which inherently require tactile interaction for accurate assessment. While integrating tactile sensing could enhance predictions, it remains impractical in this setting [5]. AttributeNet, like 1D-CNN, employs ResNet-50 for feature extraction, whereas VisionNet utilized AlexNet,

TABLE I
The MAE (%) of AttributeNet and existing methods.

Method	Real Images		Real + SD Images	
	R-S	H-S	R-S	H-S
Visual-CNN [34]	21.12	26.32	31.38	27.65
VisionNet [35]	16.74	13.76	15.41	12.86
1D-CNN [33]	14.5	10.91	9.86	11.2
AttributeNet (ours)	**9.87**	**9.21**	**7.47**	**8.16**

confirming the effectiveness of pre-trained models in improving generalization [33], [35]. Notably, the Just Noticeable Difference (JND) for MAE in texture perceptual similarity is around 10% [5], and both R-S and H-S fall below this threshold, reinforcing the proposed model's reliability in estimating haptic attributes. However, larger errors are observed for artificial materials like S7 (roof tile) and S23 (rubber mesh) due to variations in SD-images. Broad terms such as 'roof tile' or 'rubber mesh' yield diverse outputs, ranging from smooth to rough surfaces. More precise prompts, like 'fine rubber mesh with small perforations,' could improve alignment between generated and real textures. Importantly, we observed that these discrepancies can be mitigated using the attribute slider (see Sec. II-D) as discussed in the following section.

B. User Study

This study evaluated the system's usability, comfort, and quality in generating haptic and visual textures from text input.

1) Stimuli: Participants provided text prompts describing the textures they wished to generate. Haptic textures were synthesized using the pipeline described in Section II. The study included three conditions: Txt-T (Text-Texture), Txt-TI (Text-TextureImage), and Txt-TIS (Text-TextureImageSlider). In Txt-T, participants experienced haptic feedback without visual input, interacting with a blank tablet interface. In Txt-TI, haptic textures were presented with visual representations generated by the Stable Diffusion model. In Txt-TIS, participants used two interactive sliders to adjust perceptual attributes, as detailed in Sec. II-D. These conditions enabled participants to compare different levels of feedback and control.

2) Procedure: Participants sat at a table with a tablet PC and wore headphones to minimize distractions (see Fig. 3(a)). They were provided with a manual containing sample text prompts and corresponding generated images to help them understand how specific text maps to texture-based visuals and haptic feedback; examples of these prompt–image pairs are illustrated in Fig. 1. Afterwards, participants had a short practice session where they entered prompts, viewed the generated textures, and felt the haptic output to become familiar with the system before the experiment began (see Fig. 3(b)).

Measure	Description
Correctness	The generated textures accurately matched my input.
Realism	The textures felt natural and resembled real tactile sensations.
Immersion	I felt fully absorbed in the experience.
Engagement	The interaction was stimulating and held my attention.
Learning Curve	The system was intuitive and easy to learn.

Evaluation Measures Descriptions

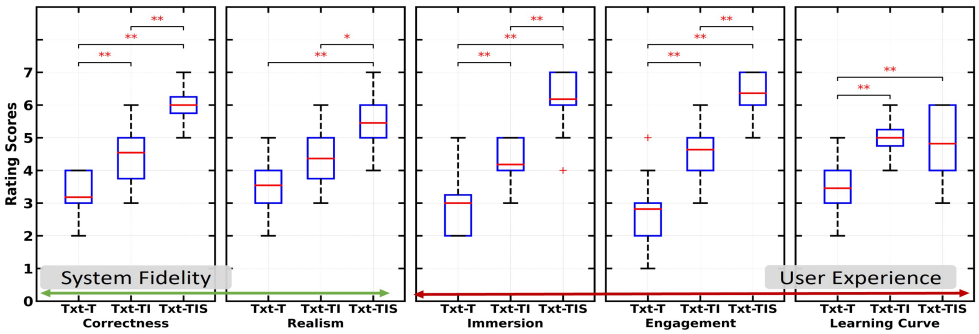


Fig. 5. Box plot of user ratings from the psychophysical study of the overall framework (right) with evaluation measure descriptions (left). Ratings were on a seven-point Likert scale. Red lines show mean ratings, and significant differences are marked with brackets and asterisks ($*p < 0.05$, $**p < 0.01$).

During the study, participants entered text prompts to generate textures and experienced them under all three conditions (Txt-T, Txt-TI, Txt-TIS). After testing each condition for a prompt, they rated the system on a seven-point Likert scale, assessing System Fidelity (realism and response accuracy) and User Experience (immersion, ease of use, and learning curve). Descriptions of these evaluation measures are provided in Fig. 5. This process was repeated for five prompts per participant, with the order of conditions randomized to prevent bias. Eleven participants (2 females, 9 males, aged 22–37, mean age: 28.7) with no reported disabilities took part in the study. Each session lasted approximately 40 minutes, and participants were compensated for their time.

3) Results and Analysis: Each participant provided 75 ratings (5 prompts $\times$ 3 conditions $\times$ 5 measures), averaged across prompts to yield 15 ratings per participant, resulting in 165 total ratings. Assumption checks confirmed homogeneity of variance (Levene's test, all $p > 0.26$), with minor deviations from normality in a few groups (e.g., Txt-TIS in Engagement: $W = 0.784$, $p = 0.006$). Given ANOVA's robustness to such violations, a one-way ANOVA was applied and confirmed statistically significant differences between conditions for all measures ($p < 0.05$). Tukey HSD post-hoc tests were used for pairwise comparisons. The mean ratings followed a consistent trend: Txt-TIS received the highest scores (5.76), Txt-TI was rated moderately higher (4.55), and Txt-T scored the lowest (3.20). Fig. 5 presents box plots illustrating the rating distributions, means, and significant differences.

Across all measures, Txt-TIS significantly outperformed the other conditions, followed by Txt-TI, with Txt-T consistently rated lowest. Correctness improved with the addition of visual feedback and interactive control ($p < 0.01$), suggesting that users found textures more accurate when they could adjust perceptual attributes. Realism was significantly higher for Txt-TIS compared to both Txt-T ($p < 0.01$) and Txt-TI ($p < 0.05$), indicating that interactive controls play a key role in making textures feel authentic. Immersion and engagement followed the same pattern ($p < 0.01$), reinforcing the idea that combining haptic, visual, and interactive elements enhances user involvement. The learning curve showed significant improvement from Txt-T to Txt-TI ($p < 0.01$) and Txt-TIS ($p < 0.05$), but no difference between Txt-TI and Txt-TIS ($p = 0.881$), suggesting interactivity does not add cognitive load. These findings confirm that combining haptic textures with visual and interactive controls results in a more accurate and engaging experience over haptic feedback alone [33].

C. Limitations and Future Work

While the proposed framework successfully integrates text-driven generative AI with haptic texture synthesis, several limitations must be considered. The system relies on Stable Diffusion for visual texture generation, which, despite its effectiveness, may introduce inconsistencies in perceptual accuracy due to variations in the generated images (see Sec. III-A1). These variations directly affect the haptic attribute prediction model, as it depends on the generated texture images. One possible improvement is fine-tuning SD with the available dataset to condition the generation process, ensuring that the output images align with the dataset's characteristics [31]. Alternatively, an ensemble approach incorporating semantic weighting from text prompts could dynamically adjust predicted attributes, enhancing robustness. Both strategies involve fine-tuning SD and leveraging text-driven semantic weighting, and they can be explored as future improvements. The framework also relies on an interpolation-based authoring algorithm with a limited texture dataset, constraining the diversity of synthesized haptic feedback. Expanding this dataset would increase the resolution of the authoring space, enabling finer variations and improving feedback accuracy. Additionally, the system maps textures using only two perceptual dimensions, rough-smooth and hard-soft, whereas real-world surfaces exhibit more complex characteristics [5], [38]. Introducing additional dimensions, such as flat-bumpy, could improve realism. Nonetheless, future perceptual studies could explore the effect of using fixed prompts across participants to gauge system reliability.

IV. Conclusion

This study introduced a novel framework for generating visual and haptic textures from text descriptions, combining Stable Diffusion for image synthesis with an interpolation-based haptic authoring approach. By linking generated images with perceptually mapped tactile signals, the system enables intuitive texture creation that aligns both visually and haptically with user input. A user study showed that incorporating visual feedback and interactive controls enhances the perceived accuracy and realism of generated textures compared to haptic-only methods. These results highlight the potential of text-driven multimodal texture generation for immersive and intuitive interactions in digital environments.

REFERENCES

[1] Y. Yoo, J. Lee, J. Seo, E. Lee, J. Lee, Y. Bae, D. Jung, and S. Choi, "Large-scale survey on adjectival representation of vibrotactile stimuli," in *Proc. HAPTICS*. New York City, United States: IEEE, 2016, pp. 393–395.

[2] K. Yoshida *et al.*, "The dimensions of tactile perception of surfaces," *Journal of Texture Studies*, vol. 12, pp. 123–135, 1968.

[3] N. Heravi, H. Culbertson, A. M. Okamura, and J. Bohg, "Development and evaluation of a learning-based model for real-time haptic texture rendering," *IEEE Transactions on Haptics*, 2024.

[4] L. R. Manfredi, H. P. Saal, K. J. Brown, M. C. Zielinski, J. F. Dammann III, V. S. Polashock, and S. J. Bensmaia, "Natural scenes in tactile texture," *Journal of neurophysiology*, vol. 111, no. 9, pp. 1792–1802, 2014.

[5] M. I. Awan, W. Hassan, and S. Jeon, "Predicting perceptual haptic attributes of textured surface from tactile data based on deep cnn-lstm network," in *Proceedings of the 29th ACM Symposium on Virtual Reality Software and Technology*, 2023, pp. 1–9.

[6] M. C. Lin and M. Otaduy, *Haptic rendering: foundations, algorithms, and applications*. CRC press, 2008.

[7] S. Shin and S. Choi, "Hybrid framework for haptic texture modeling and rendering," *IEEE Access*, vol. 8, pp. 149 825–149 840, 2020.

[8] M. I. Awan and S. Jeon, "Surface texture classification based on transformer network," *Proceedings of the Korea Human-Computer Interaction Conference (K-HCI)*, pp. 762–764, 2023.

[9] H. Culbertson, J. J. L. Delgado, and K. J. Kuchenbecker, "One hundred data-driven haptic texture models and open-source methods for rendering on 3d objects," in *2014 IEEE haptics symposium (HAPTICS)*. IEEE, 2014, pp. 319–325.

[10] A. Abdulali and S. Jeon, "Data-driven modeling of anisotropic haptic textures: Data segmentation and interpolation," in *Haptics: Perception, Devices, Control, and Applications: 10th International Conference, EuroHaptics 2016, London, UK, July 4-7, 2016, Proceedings, Part II 10*. Springer, 2016, pp. 228–239.

[11] M. I. Awan, T. Ogay, W. Hassan, D. Ko, S. Kang, and S. Jeon, "Model-mediated teleoperation for remote haptic texture sharing: Initial study of online texture modeling and rendering," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023.

[12] J. M. Romano and K. J. Kuchenbecker, "Creating realistic virtual textures from contact acceleration data," *IEEE Transactions on haptics*, vol. 5, no. 2, pp. 109–119, 2011.

[13] L. Tao, F. Wang, Y. Li, J. Wu, X. Jiang, and Q. Xi, "A cross-texture haptic model based on tactile feature fusion," *Multimedia Systems*, vol. 30, no. 3, pp. 1–12, 2024.

[14] M. I. Awan and S. Jeon, "Design and evaluation of lightweight deep learning models for synthesizing haptic surface textures," *Proceedings of the Korean Institute of Information Scientists and Engineers Conference (KIISE)*, pp. 1427–1429, 2023.

[15] W. Hassan, A. Abdulali, M. Abdullah, S. C. Ahn, and S. Jeon, "Towards universal haptic library: Library-based haptic texture assignment using image texture and perceptual space," *IEEE transactions on haptics*, vol. 11, no. 2, pp. 291–303, 2017.

[16] G. S. Giri, Y. Maddahi, and K. Zareinia, "An application-based review of haptics technology," *Robotics*, vol. 10, no. 1, p. 29, 2021.

[17] W. Hassan, A. Abdulali, and S. Jeon, "Authoring new haptic textures based on interpolation of real textures in affective space," *IEEE transactions on industrial electronics*, pp. 667–676, 2019.

[18] S. Lu, M. Zheng, M. C. Fontaine, S. Nikolaidis, and H. Culbertson, "Preference-driven texture modeling through interactive generation and search," *IEEE transactions on haptics*, pp. 508–520, 2022.

[19] Y. Ujitoko and Y. Ban, "Vibrotactile signal generation from texture images or attributes using generative adversarial network," in *Haptics: Science, Technology, and Applications: 11th International Conference, EuroHaptics 2018, Pisa, Italy, June 13-16, 2018, Proceedings, Part II 11*. Springer, 2018, pp. 25–36.

[20] W. Hassan, A. Abdulali, and S. Jeon, "Haptic texture authoring: A demonstration," in *Haptic Interaction: Perception, Devices and Algorithms 3*. Springer, 2019, pp. 18–20.

[21] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[23] A. Abdulali and S. Jeon, "Data-driven rendering of anisotropic haptic textures," in *Haptic Interaction: Science, Engineering and Design 2*. Springer, 2018, pp. 401–407.

[24] Y. Wang, A. Holynski, B. L. Curless, and S. M. Seitz, "Infinite texture: Text-guided high resolution diffusion texture synthesis," *arXiv preprint arXiv:2405.08210*, 2024.

[25] C. Gao, B. Jiang, X. Li, Y. Zhang, and Q. Yu, "Genesistex: Adapting image denoising diffusion to texture space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4620–4629.

[26] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[27] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[28] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International conference on machine learning*. Pmlr, 2021, pp. 8821–8831.

[29] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 278–25 294, 2022.

[30] S. Patil, P. Cuenca, N. Lambert, and P. von Platen, "Stable diffusion with diffusers," *Hugging Face Blog*, 2022, available at: https://huggingface.co/blog/stable_diffusion.

[31] M. Stroinski, K. Kwarciak, M. Kowalewski, D. Hemmerling, W. Frier, and O. Georgiou, "Text-to-haptics: Enhancing multisensory storytelling through emotionally congruent midair haptics," *Advanced Intelligent Systems*, p. 2400758, 2024.

[32] M. I. Awan and S. Jeon, "Estimating perceptual attributes of haptic textures using visuo-tactile data," *arXiv preprint arXiv:2505.16352*, 2025.

[33] W. Hassan, J. B. Joolee, and S. Jeon, "Establishing haptic texture attribute space and predicting haptic attributes from image features using 1d-cnn," *Scientific Reports*, vol. 13, no. 1, p. 11684, 2023.

[34] Y. Gao, L. A. Hendricks, K. J. Kuchenbecker, and T. Darrell, "Deep learning for tactile understanding from visual and haptic data," in *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 536–543.

[35] H. Zheng, L. Fang, M. Ji, M. Strese, Y. Özer, and E. Steinbach, "Deep learning for surface material classification using haptic and visual information," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2407–2416, 2016.

[36] S. Tsuji and T. Kohama, "Using a convolutional neural network to construct a pen-type tactile sensor system for roughness recognition," *Sensors and Actuators A: Physical*, vol. 291, pp. 7–12, 2019.

[37] Z. Lin, H. Zheng, Y. Lu, J. Zhang, G. Chai, and G. Zuo, "Object surface roughness/texture recognition using machine vision enables for human-machine haptic interaction," *Frontiers in Computer Science*, vol. 6, p. 1401560, 2024.

[38] W. Hassan, M. I. Awan, A. Raza, K.-U. Kyung, and S. Jeon, "Quantifying haptic affection of car door through data-driven analysis of force profile," *arXiv preprint arXiv:2411.11382*, 2024.