Generating Tactile Textures from Perceptual Descriptors with Diffusion Models: A Feasibility Study

Quinn Begelinger
Department of Cognitive Robotics
Delft University of Technology
Delft, Netherlands
quinn.begelinger@gmail.com

Yasemin Vardar
Department of Cognitive Robotics
Delft University of Technology
Delft, Netherlands
y.vardar@tudelft.nl

Abstract

Capturing high-quality tactile signals typically requires specialized hardware and controlled laboratory conditions, limiting the scalability and diversity of haptic content. Generative models, which have transformed digital language, vision, and audio content, offer a promising alternative for haptics. We propose a two-stage latent diffusion framework for generating tactile texture signals conditioned on psychophysical descriptors. In the first stage, a diffusion model learns a compact latent representation of friction signals produced by a finger sliding over diverse surfaces and reconstructs them with high temporal fidelity. In the second stage, a diffusion-based encoder maps perceptual ratings, such as roughness, bumpiness, and slipperiness, into this latent space, enabling texture generation from perceptual input. Reconstruction results demonstrate low error and a realistic signal structure. However, conditioning on psychophysical descriptors produces limited variations, primarily affecting signal amplitude, highlighting an open challenge in perceptually conditioned generative haptics.

CCS Concepts

• **Human-centered computing**;

Keywords

Tactile Textures, Generative AI, Diffusion models

ACM Reference Format:
Quinn Begelinger and Yasemin Vardar. 2026. Generating Tactile Textures from Perceptual Descriptors with Diffusion Models: A Feasibility Study. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3772363.3799381

1 Introduction

In the modern digital landscape, interaction is dominated by visual and auditory experiences, while touch remains underrepresented despite its central role in perceiving and interacting with the physical world. Although haptic feedback is increasingly integrated into consumer devices, such as smartphones, smartwatches, and game consoles, these sensations are typically limited to simple vibrotactile cues, far from the rich tactile experiences of real objects.

CHI EA '26, Barcelona, Spain

ACM ISBN 979-8-4007-2281-3/26/04
https://doi.org/10.1145/3772363.3799381

Rendering realistic material properties, especially surface texture, has therefore become a primary focus of research, with significant progress in technologies that support direct bare-finger interaction [3]. Many approaches rely on record-and-playback [5, 6, 14], in which tactile data is captured under controlled conditions and stored in texture databases covering various materials and exploration patterns. Yet acquiring high-quality tactile signals requires specialized laboratory equipment and expertise, making it impractical to represent the full diversity of real-world materials. Also, these methods offer limited flexibility for modifying or enhancing the generated sensations. These limitations continue to constrain the scalability, richness, and broader adoption of haptic technologies.

To address these challenges, recent studies have explored the synthesis of tactile texture signals using either signal feature interpolation [7, 11] or image-based generation methods [4, 19]. Interpolation-based methods extract handcrafted signal features, such as Mel-frequency cepstral coefficients or autoregressive model parameters, from recorded signals and synthesize new ones by interpolating them based on an affective or perceptual space derived from subjective evaluations or user preferences using models like Generative Adversarial Networks (GANs). Image-based approaches employ similar models to infer tactile signals from visual surface information, sometimes combined with recorded contact data [8, 19].

While these approaches have demonstrated promising results, they suffer from several limitations. Interpolation-based approaches are constrained by the coverage of the original dataset and rely on predefined signal features, limiting their ability to generalize beyond observed textures. Image-based methods infer tactile properties indirectly from visual cues, which do not reliably encode frictional or vibrational characteristics and do not explicitly model user interaction dynamics. Although perceptual information may be incorporated during training or optimization, neither class of methods provides a principled mechanism for directly controlling generated textures through explicit perceptual descriptors at inference time. Recent human-in-the-loop approaches partially address this limitation by iteratively adjusting latent representations based on user feedback [21], but such methods require repeated user interaction and do not offer immediate, parameterized control in perceptual space.

Here, we propose a psychophysical feature–to–texture diffusion framework for generating friction signals arising from finger–surface interactions, conditioned on human perceptual descriptors. The approach employs a two-stage latent diffusion architecture in which a compact latent representation of tactile textures is first learned and subsequently decoded into full-resolution friction signals. By conditioning generation on perceptual ratings, such as

rough–smooth, flat–bumpy, and sticky–slippery, the framework enables the synthesis of tactile signals guided by intended perceptual qualities. The initial results show high-fidelity reconstruction but limited variation in generated signals with perceptual conditioning. These findings indicate that diffusion-based generative models can support scalable haptic content creation without complex laboratory setups or specialized sensing hardware, while also highlighting challenges in perceptually driven texture synthesis.

2 Methods

2.1 Dataset Preparation

We used two subsets of the publicly available SENS3 dataset [1], containing finger–surface interaction data for 50 distinct surfaces and their corresponding psychophysical ratings.

The first subset comprises finger-on-texture ***sliding interactions*** recorded during the free exploration of 50 surfaces by two participants, including 3D force data (with a sampling rate of 10 kHz) and finger velocity. As texture signals, we extracted the contact forces in the sliding direction [13]. Due to the unconstrained nature of the recordings, data lengths vary per texture, and both normal force and finger velocity fluctuate widely. Hence, we selected texture signal segments when the normal force was between 0.4–0.6 N and velocity between 66–99 mm/s, corresponding to comfortable interaction parameters for friction-modulation displays [2]. The minimum segment length was chosen as 1.64 s and brief deviations in the normal force and velocity for a maximum of 0.1 s were allowed to maintain continuity. We applied a 50% overlap between consecutive segments to increase dataset size and ensure temporal features truncated at segment edges are captured centrally in other segments. After slicing and overlapping, the dataset contained approximately 22,000 texture samples for training. A bandpass butterworth filter was applied to the extracted samples to remove frequencies below 20 Hz and above 1 kHz to eliminate low-frequency finger motion and high-frequency signal content above the perceivable range [13].

The second subset consists of ***psychophysical ratings*** from 12 participants for each of the 50 textures. While the SENS3 dataset provides ratings along eight adjective axes, we focused on three most relevant to friction: rough–smooth, slippery–sticky, and bumpy–even. Each axis was rated independently on a 9-point scale (0–8). To standardize the data, we z-score normalized each participant's ratings for each axis, ensuring a standard normal distribution. This procedure balanced contributions across participants and mitigated individual biases, resulting in mean ratings of 0, with most values ranging between -2 and 2.

2.2 Modeling Framework and Implementation

For texture synthesis, we employed diffusion models, which operate through two complementary processes: a forward process, progressively adding noise to the input data, and a reverse process, iteratively removing noise to reconstruct the original sample [12, 15]. In the forward process, textures are gradually corrupted over a series of discrete diffusion steps, with the amount of noise controlled by a predefined schedule. The reverse process uses a neural network to iteratively remove this noise, producing high-fidelity reconstructions. This approach enables flexible generation with relatively few denoising steps while preserving both global structure and fine temporal detail in the synthesized textures. We chose diffusion models for their superior performance in producing high-quality, diverse images [17] and music [9].

Our framework uses two diffusion models working in series with each other; see Figure 1a for the full inference pipeline:

1. LCTG (Latent-Conditioned Texture Generator): This model converts a texture signal into a compact latent representation and then reconstructs it. The latent representation is a learned, low-dimensional encoding of the texture's perceptually salient spectro-temporal structure. Rather than preserving raw waveform detail, it captures global temporal patterns and frequency content that are relevant to tactile perception, while discarding fine-grained variations that are less perceptually meaningful. The latent space is learned jointly with a diffusion-based decoder, making it approximately invertible: latent codes retain sufficient information to reconstruct realistic texture signals while remaining abstract enough to support perceptually driven manipulation and generation.

The process begins by transforming the raw signal into a Mel-spectrogram, a two-dimensional time–frequency representation. The Mel scale emphasizes lower frequencies, which are particularly salient in tactile perception. A 10-layer 1D convolutional encoder then compresses the spectrogram along the temporal axis, summarizing the texture into a latent vector while preserving essential temporal patterns. This compression substantially reduces computational load and allows subsequent stages to focus on the most informative texture features.

The latent vector is passed to a 1D U-Net–based diffusion decoder [20] with seven downsampling layers and seven corresponding upsampling layers, which reconstructs the texture in the time domain. The downsampling path captures coarse, global structure, while the upsampling path restores fine temporal detail. To prevent the loss of subtle texture features during compression, skip connections directly transfer intermediate representations from earlier stages to later reconstruction stages. This structure allows fine-grained temporal information to bypass the compression bottleneck and be reintroduced during reconstruction. In practice, it enables the model to combine global texture characteristics with local, high-frequency details that are essential for generating realistic haptic textures. The decoder is trained using a v-objective diffusion loss [16], which improves training stability and enables high-quality reconstruction with fewer denoising steps. An optional frequency-aware loss term was also included to compare the spectral content of the generated and target signals, encouraging perceptual consistency in the frequency domain. During inference (the process of generating a texture signal from a set of input ratings using the trained DALE–LCTG pipeline), we employed DDIM sampling [18], which significantly reduces the number of required denoising steps compared to standard diffusion sampling, making the LCTG suitable for real-time or interactive texture generation.

2. DALE (Diffusion Adjective-to-Latent Encoder): This model maps psychophysical ratings into the latent space learned by the LCTG encoder. It therefore operates in a perceptually structured latent space, in which variations correspond to meaningful changes in texture characteristics as captured by the encoder, rather than to low-level signal differences.

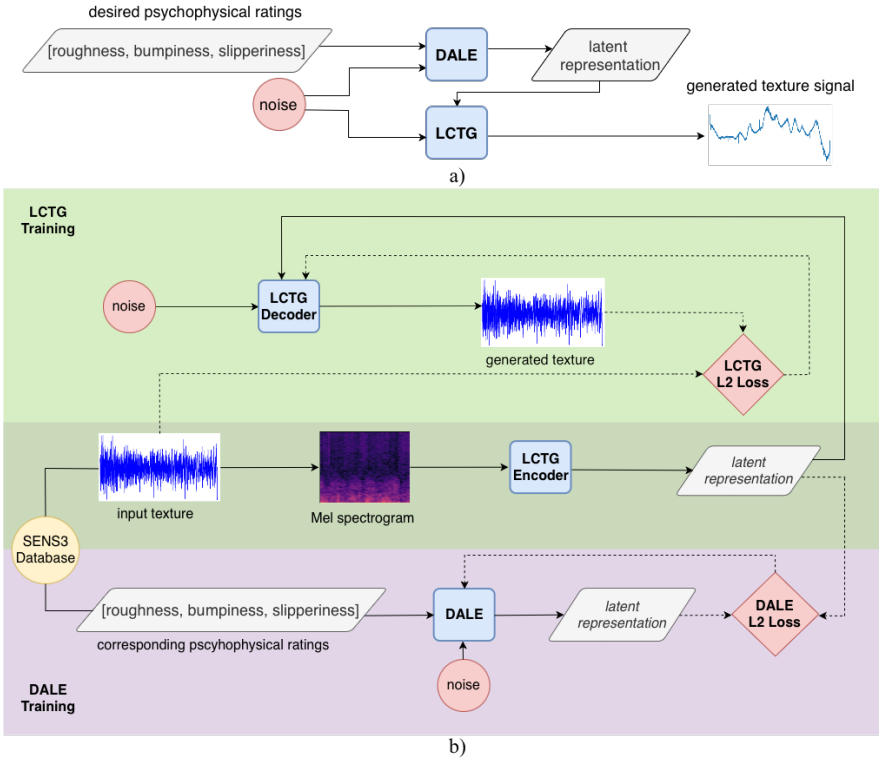

Figure 1: Overview of the proposed framework: (a) two-stage inference process and (b) training pipeline. Solid arrows indicate the forward flow of input and output data through the models during training, while dotted arrows denote data paths used for loss computation and backpropagation. The training procedures of LCTG and DALE are highlighted in green and purple, respectively. The overlapping green region indicates that the LCTG encoder is trained as part of LCTG training and subsequently reused to compute latent representations for DALE training. Noise nodes represent input signals corrupted with additive noise as required by the diffusion process.

DALE uses the same 1D U-Net backbone and diffusion formulation as the LCTG decoder, but operates entirely in the latent space rather than on time-domain signals. Psychophysical ratings are embedded into a higher-dimensional representation using sinusoidal projections and integrated into the network via cross-attention layers applied at every U-Net resolution level, with learnable context scaling. This mechanism allows the model to modulate different aspects of the latent representation according to perceptual input. To ensure that the model does not ignore the ratings, an auxiliary conditioning loss encourages active use of the psychophysical information. As a result, DALE learns to generate novel latent representations aligned with human perception, rather than merely reproducing latent codes corresponding to existing textures. This two-stage design allows DALE to handle the computationally intensive perceptual conditioning upfront, keeping the LCTG lightweight and suitable for real-time texture generation.

2.3 Training

We adopted a two-phase training strategy (Figure 1b). First, the LCTG was trained to reconstruct texture signals. It initially learned generalizable texture features from the full dataset and was then fine-tuned on a subset recorded under controlled finger speed and force conditions. This two-step strategy improves reconstruction fidelity and ensures that the latent representation captures consistent, perceptually relevant features. Next, DALE was trained on the same filtered subset. Latent vectors extracted using the pretrained LCTG encoder were paired with psychophysical ratings, allowing the model to learn how human perception maps into the latent space. This way, DALE can generate new latent representations corresponding to novel ratings, which can subsequently be decoded into realistic texture signals.

3 Results

3.1 Diffusion Step Efficiency

During inference, the number of diffusion steps can be selected to trade off reconstruction quality against computational cost. While higher step counts generally improve accuracy, the v-objective diffusion formulation enables high-quality results even with relatively few steps. On an NVIDIA RTX 4060, over 50 steps were performed in real-time, yielding a waveform RMSE (root mean squared error between the original and generated time-domain signals) as low as 0.0060 N. Evaluating the efficiency score—product of RMSE and inference time—identified 10 diffusion steps as the optimal balance of

speed and accuracy and used this value to calculate all subsequent results.

3.2 LCTG Reconstruction Performance

The Latent-Conditioned Texture Generator was evaluated by reconstructing textures in the validation set and comparing them to the originals. Time-domain analysis revealed that the reconstructed signals closely matched the originals in magnitude and temporal structure, with key low-frequency components largely preserved. However, frequency-domain analysis revealed systematic deviations: low frequencies were sometimes underrepresented, and a relatively larger high-frequency noise was present above 1 kHz. Quantitatively, the mean waveform RMSE was 0.0114 N (median 0.0107, range 0.0062–0.0181). Signal power was well preserved (mean power difference 5.0×10^{-5} N^2, max 1.0×10^{-4}), with an average signal to noise ratio of −1.87 dB (range −2.36 to −1.21 dB). These negative values primarily reflect the SNR computation relative to zero-mean signal power. Removing the DC component constrains signal power to the variance of temporal fluctuations; as a result, even minor reconstruction errors (e.g., amplified high-frequency noise) can produce a negative decibel ratio despite low absolute RMSE. To our knowledge, there is no established reference range for "good" SNR in haptic texture generation, limiting direct comparison with prior work.

3.3 DALE Latent Faithfulness

To assess DALE's sensitivity to conditioning inputs, we measured the *latent shift magnitude*, defined as the L2 distance between latent vectors generated from varying psychophysical ratings. With the noise vector fixed, each adjective (Roughness, Bumpiness, Slipperiness) was varied independently in increments of 0.2, while others were held at zero; 30 latent vectors per rating were generated and compared to the baseline. The results showed that the normalized shift magnitudes (Figure 2), were small (0–2%). Roughness and Bumpiness showed weak but consistent increases as ratings diverge from the mean, suggesting modest responsiveness, while Slipperiness showed no systematic trend, indicating limited sensitivity along this dimension.

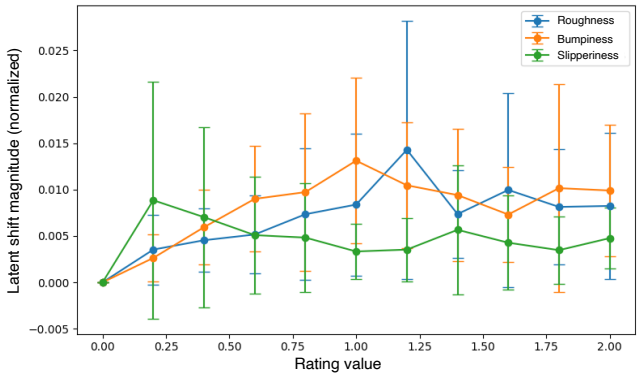

Figure 2: Magnitude of latent-space shifts (normalized) as a function of perceptual rating values, used to assess the sensitivity of DALE to psychophysical descriptors.

3.4 Full Pipeline Inference

The complete DALE–LCTG pipeline was evaluated by generating textures from participant ratings and comparing them to original validation segments. Across 20 randomly selected segments with five repetitions per segment, each using a different randomly sampled set of ratings, the mean waveform RMSE was 0.00557 ± 0.00109 N.

To assess the effect of individual perceptual rating dimensions, we varied one rating at a time from −2 to +2 standard deviations (in 0.5 increments), while sampling the other two ratings independently from a standard normal distribution. For each fixed rating, 50 textures were generated, and their average frequency spectra were computed using the Fast Fourier Transform. Across all conditions, frequency peaks remained largely stationary, with variations primarily in amplitude rather than spectral shape. Smoothness and Slipperiness showed gain changes mainly at low frequencies (<200 Hz), with Smoothness also exhibiting effects at higher frequencies (>200 Hz). Overall, all three perceptual dimensions had comparable influence, and the averaged spectra suggest limited frequency diversity in the generated textures.

We also evaluated the similarity between real and generated textures at the distribution level using the Frechet Audio Distance (FAD) [10], a metric commonly used in audio generation. FAD measures how closely the overall statistical distribution of generated signals matches that of real signals, considering features such as spectral patterns. Using 500 textures generated from randomly sampled ratings, the FAD between the generated set and the full database of real textures was 21.62. For context, state-of-the-art music generation models [16] typically achieve FAD scores below 5. This indicates that, while the generated textures are perceptually plausible, their overall spectral statistics remain more distant from real textures compared to highly refined audio generation models.

We note that the absolute values of FAD should be interpreted with caution in the context of haptic texture generation. This metric was originally developed for audio quality assessment and does not have well-established reference ranges or "acceptable" thresholds in the haptics domain. While experiments with human participants would provide the most reliable assessment of perceptual quality, we did not include such an evaluation in this study because the limited variation observed under perceptual conditioning suggests that further improvements in the model and training data are needed before such an evaluation would be informative.

4 Discussion

We presented a two-stage latent diffusion framework for generating tactile texture signals conditioned on psychophysical descriptors. By explicitly decoupling texture reconstruction from perceptual conditioning, the proposed architecture separates the learning of low-level signal structure from the mapping between human perception and latent representations. This design enables the model to translate subjective descriptors—such as roughness, bumpiness, and slipperiness—into friction signals while maintaining computational efficiency.

Our results demonstrate that diffusion models are well-suited for modeling tactile texture signals. The Latent-Conditioned Texture Generator (LCTG) achieves low reconstruction error. and reliably

preserves the dominant temporal structure and overall energy of real textures, indicating that diffusion-based generative models can capture the stochastic and non-periodic characteristics of friction-based tactile data. However, reconstruction fidelity in the frequency domain remains limited. In particular, low-frequency components that are perceptually salient are sometimes underrepresented, while spurious high-frequency energy occasionally appears despite being absent from the training data. These artifacts are perceptually undesirable and suggest a bottleneck in the latent representation or loss formulation. Potential remedies include increasing latent capacity, introducing perceptually weighted or frequency-filtered loss functions tailored to frictional tactile data (rather than the Mel scale), and improving the spectral balance of the training data using recordings from more participants and textures.

We further observe that variations in psychophysical ratings lead to measurable but modest changes in the generated textures. Analysis of latent representations shows only small shifts in response to changes in individual rating dimensions, indicating that the conditioning signal influences generation but does not dominate it. Consistently, frequency-domain analysis reveals that conditioning primarily affects the overall amplitude of the signal rather than inducing substantial changes in spectral shape. Moreover, the influence of different perceptual dimensions is qualitatively similar, suggesting limited disentanglement between axes such as roughness, bumpiness, and slipperiness. These limitations can largely be attributed to characteristics of the dataset, including correlated perceptual dimensions, uneven rating distributions, and substantial inter-participant disagreement [1]. In addition, texture signals and perceptual ratings were collected from different participant groups, weakening the correspondence between signal properties and subjective labels and further constraining the learnable conditioning relationship.

Overall, this work demonstrates the feasibility of using diffusion models to generate tactile signals, while also highlighting the challenges of synthesizing signals conditioned on psychophysical descriptors. Given the limited variation observed under perceptual conditioning, we did not include a human evaluation of the generated outputs in this study. Future work should focus on collecting more balanced and generalizable datasets, strengthening the correspondence between tactile signals and perceptual ratings, exploring perceptually motivated loss functions, and validating generated textures through human evaluation.

Acknowledgments

This study was partially funded by the Dutch Research Council (NWO) under the VENI scheme (project number 19153) and the European Research Council (ERC) under the Starting Grant scheme (project number 101220242). The authors also thank Jagan K. Balasubramanian for his assistance with the SENS3 dataset.

References

[1] Jagan K. Balasubramanian, Bence L. Kodak, and Yasemin Vardar. 2025. SENS3: Multisensory Database of Finger-Surface Interactions and Corresponding Sensations. In *Haptics: Understanding Touch; Technology and Systems; Applications and Interaction*, Hiroyuki Kajimoto et al. (Eds.). Springer Nature Switzerland, Cham, 262–277.

[2] Jagan K. Balasubramanian, Daan M. Pool, and Yasemin Vardar. 2026. Sliding speed influences electrovibration-induced finger friction dynamics on touchscreens. *Tribology International* 213 (2026), 111054. doi:10.1016/j.triboint.2025.111054

[3] Cagatay Basdogan, Frederic Giraud, Vincent Levesque, and Seungmoon Choi. 2020. A Review of Surface Haptics: Enabling Tactile Effects on Touch Surfaces. *IEEE Transactions on Haptics* 13, 3 (2020), 450–470.

[4] Shaoyu Cai, Lu Zhao, Yuki Ban, Takuji Narumi, Yue Liu, and Kening Zhu. 2022. GAN-based image-to-friction generation for tactile simulation of fabric material. *Computers & Graphics* 102 (2022), 460–473.

[5] Tamara Fiedler and Yasemin Vardar. 2019. A Novel Texture Rendering Approach for Electrostatic Displays. In *International Workshop on Haptic and Audio Interaction Design - HAID2019*. Lille, France. https://hal.science/hal-02011782

[6] Roman V Grigorii, Roberta L Klatzky, and J Edward Colgate. 2021. Data-driven playback of natural tactile texture via broadband friction modulation. *IEEE Transactions on Haptics* 15, 2 (2021), 429–440.

[7] Waseem Hassan, Arsen Abdulali, and Seokhee Jeon. 2020. Authoring New Haptic Textures Based on Interpolation of Real Textures in Affective Space. *IEEE Transactions on Industrial Electronics* 67, 1 (2020), 667–676.

[8] Negin Heravi, Heather Culbertson, Allison M. Okamura, and Jeannette Bohg. 2024. Development and Evaluation of a Learning-Based Model for Real-Time Haptic Texture Rendering. *IEEE Transactions on Haptics* 17, 4 (2024), 705–716.

[9] Qingqing Huang, Daniel S. Park, Tao Wang, Timo I. Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, Jesse Engel, Quoc V. Le, William Chan, Zhifeng Chen, and Wei Han. 2023. Noise2Music: Text-conditioned Music Generation with Diffusion Models. arXiv:2302.03917 [cs.SD]

[10] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2019. Fréchet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms. arXiv:1812.08466 [eess.AS]

[11] Shihan Lu, Mianlun Zheng, Matthew C. Fontaine, Stefanos Nikolaidis, and Heather Culbertson. 2022. Preference-Driven Texture Modeling Through Interactive Generation and Search. *IEEE Transactions on Haptics* 15, 3 (2022), 508–520.

[12] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. 2022. Diffusion Autoencoders: Toward a Meaningful and Decodable Representation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10609–10619.

[13] Benjamin A Richardson, Yasemin Vardar, Christian Wallraven, and Katherine J Kuchenbecker. 2022. Learning to feel textures: Predicting perceptual similarities from unconstrained finger-surface interactions. *IEEE Transactions on Haptics* 15, 4 (2022), 705–717.

[14] Joseph M. Romano and Katherine J. Kuchenbecker. 2012. Creating Realistic Virtual Textures from Contact Acceleration Data. *IEEE Transactions on Haptics* 5, 2 (2012), 109–119.

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10674–10685.

[16] Flavio Schneider, Ojasv Kamal, Zhijing Jin, and Bernhard Schölkopf. 2024. Moûsai: Efficient Text-to-Music Diffusion Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 8050–8068.

[17] Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. 2020. Improving Image Captioning with Better Use of Captions. arXiv:2006.11807 [cs.CV]

[18] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2022. Denoising Diffusion Implicit Models. arXiv:2010.02502 [cs.LG]

[19] Yusuke Ujitoko and Yuki Ban. 2018. Vibrotactile Signal Generation from Texture Images or Attributes Using Generative Adversarial Network. In *Haptics: Science, Technology, and Applications*, Domenico Prattichizzo et al. (Eds.). Springer International Publishing, Cham, 25–36.

[20] Xue Xia, Daiwei Zhang, Wenxuan Song, Wei Huang, and Lorenz Hurni. 2025. MapSAM: Adapting segment anything model for automated feature detection in historical maps. *GIScience & Remote Sensing* 62, 1 (2025), 2494883.

[21] Mingxin Zhang, Shun Terui, Yasutoshi Makino, and Hiroyuki Shinoda. 2025. TexSenseGAN: A User-Guided System for Optimizing Texture-Related Vibrotactile Feedback Using Generative Adversarial Network. *IEEE Transactions on Haptics* 18, 2 (2025), 325–339.