AudioHapticDiffusion: Synchronized Audio-Haptic Generation via Shared Latent Space

Arata Kotani*[1][0009−0001−7421−5157] and Arata Horie*[1][0000−0002−8161−5036]

commissure, Inc., Tokyo, Japan
{a.kotani,a.horie}@commissure.co.jp
https://commissure.co.jp

Abstract. This study proposes a dual-branch latent diffusion model that simultaneously generates sound effects and a single-channel vibrotactile waveform (for wideband actuators in handheld/wearable devices) from text prompts. By sharing the latent space of a pretrained audio generation model, the method inherently ensures temporal synchronization between audio and vibration. The audio branch uses a frozen diffusion model while the vibration branch employs a Transformer-CNN decoder, further refined with Direct Preference Optimization (DPO). Evaluation with 10 participants demonstrated that the proposed method achieves performance equal to or better than rule-based audio-to-vibration transformations (e.g., low-pass filtering, envelope modulation), with notably lower distraction scores.

Keywords: Haptic Generation · Latent Diffusion Model · Audio-Haptic Synchronization · Text-to-Haptic · Multimodal Generation

1 Introduction

Real-world experiences are formed through the temporal synchronization of multiple sensory modalities. For example, when typing on a keyboard, the visual information of a finger touching a key, the auditory information of the keystroke sound, and the tactile information of the impact transmitted to the fingertip synchronize within tens of milliseconds to form a coherent perceptual experience of "pressing a key" [14]. Such multisensory integration mechanisms form the foundation of the human perceptual system, and in immersive experience technologies such as virtual reality (VR) and augmented reality (AR), appropriately synchronizing haptic feedback with visual and auditory information is key to immersion and realism [13]. Perceptual psychology research indicates that for synchrony perception between auditory and tactile modalities to be established, the time difference between them must be approximately 55 ms or less [8]; exceeding this threshold causes both stimuli to be perceived as separate events, compromising experiential coherence.

* These authors contributed equally to this work.

In recent years, generative models based on deep learning have made remarkable progress across various modalities including text, images, and audio. In particular, diffusion models [1] have achieved high-quality outputs in image generation [12] and audio generation [2, 7], and text-to-audio methods represented by Make-An-Audio 2 and AudioLDM have enabled intuitive content creation through natural language. Meanwhile, in the field of haptic generation, TactGAN [15] proposed a GAN-based method for generating tactile textures from images, and HapticGen [5] demonstrated a method for directly generating vibration patterns from text using a Transformer architecture. More recently, Scene2Hap [3] coupled LLM-based scene reasoning with physical modeling to author vibrotactile signals for full VR scenes, and Sound2Hap [6] learned an audio-to-vibrotactile mapping supervised by human ratings. Both, however, assume that a VR scene description or an audio waveform is already available, so cross-modal alignment is handled post-hoc; in contrast, our method generates audio and vibration jointly from a single text prompt through a shared latent space, making synchrony a structural property of the architecture. Throughout this paper, *rule-based methods* refers to deterministic audio-to-vibration transformations—typically low-pass filtering [10], envelope-based carrier modulation [9], or Intensity Segment Modulation (ISM) [16]—which are lightweight but cannot adapt vibration to the semantic category of the audio event.

This study proposes a dual-branch latent diffusion model that simultaneously generates sound effects and vibrotactile feedback from text prompts, targeting single-channel waveforms for wideband actuators in handheld/wearable devices (texture rendering, kinesthetic feedback, and multi-point spatial haptics are out of scope). The core of the proposed method lies in sharing the latent space in the generation process of audio and vibration. Using a pretrained text-conditioned audio diffusion model as the foundation, an audio decoder branch that generates audio spectrograms from the latent space and a haptic decoder branch that generates vibration spectrograms from the same latent representation are arranged in parallel. Since both branches take the common latent representation as input, the generated audio and vibration are inherently guaranteed to be temporally synchronized. Furthermore, we introduce optimization based on human preferences using Direct Preference Optimization (DPO) [11] to improve the subjective quality of generated vibrations.

The contributions of this study are as follows.

- We propose a method that inherently guarantees temporal synchrony through a dual-branch architecture that shares the latent space for audio and vibration generation.
- We quantitatively evaluated the temporal synchrony of generated audio and vibration using objective metrics of onset detection and envelope correlation.
- We conducted an objective evaluation on 12 audio scenes and a user study on 9 of them to demonstrate the effectiveness of the proposed method through comparison with rule-based methods.

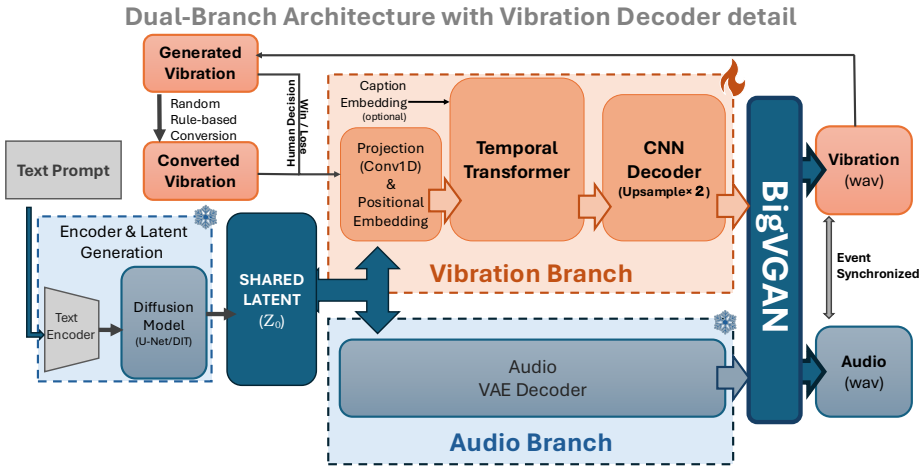

Fig. 1. Dual-branch architecture for synchronized audio-haptic generation. Snowflake icons indicate frozen pretrained components; flame icons indicate trainable components.

2 Method

2.1 System Overview

The proposed system is a dual-branch latent diffusion model that simultaneously generates sound effects and vibrotactile feedback from text prompts. It is built upon a pretrained text-conditioned audio diffusion model (Make-An-Audio 2) and adopts a configuration that decodes audio and vibration in parallel from its latent space. Make-An-Audio 2 was chosen because its spectrogram-aligned latent preserves the onset-level temporal structure needed for audio–haptic synchrony and its publicly released weights and ~330,000-clip training corpus are directly reusable for vibration supervision; any diffusion-based text-to-audio backbone with analogous properties should be compatible.

The system comprises several key components working in unison (Figure 1). A **CLAP Text Encoder** first converts text prompts into semantic feature vectors (caption embedding). These vectors condition a **Diffusion Model (UNet/DiT)**, which iteratively generates latent representations from noise. These latent representations are then processed by two parallel decoders: a **VAE Decoder** that reconstructs audio spectrograms, and a specialized VibrationDecoder that generates vibration spectrograms from the same latent features. Finally, a **BigVGAN Vocoder** converts both spectrograms into their respective waveforms.

2.2 Dual-Branch Architecture

The core of the proposed method lies in generating audio and vibration from a shared latent space. The latent representation $z \in \mathbb{R}^{C \times T}$ generated by the

UNet is input to both the VAE decoder and VibrationDecoder, which output audio spectrogram $\hat{S}_a$ and vibration spectrogram $\hat{S}_v$, respectively. This design ensures that both modalities share the same temporal structure, structurally guaranteeing temporal synchronization.

The VibrationDecoder adopts a hybrid Transformer-CNN structure. The Transformer encoder projects the latent sequence, adds positional encoding, and integrates temporal context via self-attention, with cross-attention for text conditioning. The CNN decoder then upsamples the features to generate the vibration spectrogram (80 mel channels).

During training, only the VibrationDecoder is updated; the Diffusion Model, VAE, CLAP encoder, and BigVGAN vocoder remain frozen (Figure 1), which keeps training efficient and guarantees that adding the vibration branch cannot degrade the underlying text-to-audio model. Training data was prepared from the same audio dataset used to train Make-An-Audio 2 (approximately 330,000 samples). Each audio sample was first processed through Spleeter[1] to remove voice components, then converted to pseudo-vibration data through rule-based transformation. The transformation consists of two components: (1) low-frequency extraction using a 200 Hz low-pass filter, chosen to match the practical pass-band of wideband LRA / voice-coil actuators in consumer devices, and (2) envelope extraction followed by modulation of a 55 Hz sine-wave carrier, placed within the frequency range where handheld actuators such as the DualSense produce clear "knock"-like sensations. These two components are mixed at a 2:8 ratio (20% low-pass, 80% envelope modulation) and converted to mel spectrograms. The VibrationDecoder receives the shared latent representation and caption embedding as inputs, and is supervised using this converted vibration spectrogram as the target. The model is trained using the Mean Squared Error (MSE) loss between the generated and target vibration spectrograms:

$$L_{vib} = \mathbb{E}_z[\|\hat{S}_v - S_v\|_2^2] \tag{1}$$

2.3 Direct Preference Optimization (DPO)

The MSE objective in Eq. (1) is insensitive to perceptual axes such as impact sharpness or residual over-vibration, so we additionally fine-tuned the VibrationDecoder with Direct Preference Optimization (DPO) to reshape the output along these axes while leaving the shared-latent structure untouched. Throughout this paper, *subjective quality* denotes human-judged naturalness and non-intrusiveness of vibration in the context of its paired audio, as opposed to signal-level reconstruction fidelity. The DPO mechanism requires pair-wise preference data, which we constructed as follows (Figure 1).

Preference Data Construction. For each training sample, we prepare two vibration signals: Data A (model-generated) and Data B (transformed version of Data A). Transformations include frequency modifications (filtering, spectral

[1] Spleeter is an open-source audio separation library developed by Deezer: https://github.com/deezer/spleeter

adjustment), temporal modifications (envelope reshaping, gain modulation), and perturbation injection (noise, clicks). The transformed signal is blended with the original and normalized.

Human Evaluation. Human annotators are presented with both Data A and Data B in randomized order and evaluate which vibration better matches the corresponding audio. This fair comparison creates preference pairs (y_w, y_l) where y_w denotes the chosen (winning) vibration and y_l denotes the rejected (losing) one. As an initial investigation, we collected 200 preference pairs for this study.

Optimization. Using the collected preference data, the VibrationDecoder is fine-tuned with the caption embedding (derived from the original text prompt) as input. The optimization applies a correction to the base MSE loss that pushes the model output away from rejected samples toward preferred ones:

$$L_{DPO} = -\mathbb{E}[\log \sigma(\beta(r(y_w) - r(y_l)))] \tag{2}$$

Here, $r(\cdot)$ is the implicit reward for the VibrationDecoder output, and β is the temperature parameter. This mechanism enables refinement of vibration quality based on human perceptual preferences.

2.4 Inference Pipeline

During inference, the text prompt is encoded via CLAP, and a latent representation is generated using DDIM sampling. This latent is decoded in parallel by the VAE and VibrationDecoder, then converted to waveforms via BigVGAN.

3 Experiments

3.1 Quantitative Evaluation

To objectively evaluate the temporal synchrony of generated audio and vibration, we conducted a quantitative evaluation using two primary metrics: *Onset Time Difference* and *Envelope Correlation*. Onset Time Difference compares the onset detection results of audio and vibration, calculating time differences through nearest neighbor matching, with 55 ms considered the perceptual synchrony threshold [8]. Envelope Correlation calculates the Pearson correlation coefficient between the envelopes of both signals. We compared our PROPOSED dual-branch model against three baselines: a LOWPASS condition applying a 200 Hz low-pass filter to generated audio; an ENVELOPE condition modulating generated audio with a 55 Hz sine wave; and a NOISE condition using random pink noise as a baseline. We evaluated 12 types of audio scenes, including footsteps, door slams, ball bounces, glass shattering, rain, engine sounds, buzzing, scraping, cafe ambience, train station sounds, conversation, and screams with gunshots.

As shown in Table 1, the mean onset time difference of the proposed method was 54.45 ms. Regarding envelope correlation, the proposed method ($r = 0.550$) yielded values between the rule-based methods and the NOISE condition. A

Table 1. Quantitative Evaluation Results (by Condition)

Condition	Onset Time Diff. (ms)	Envelope Corr. (r)
PROPOSED	54.45 ± 37.07	0.550 ± 0.431
LOWPASS	33.52 ± 37.62	0.850 ± 0.209
ENVELOPE	27.43 ± 30.13	0.974 ± 0.072
NOISE	58.81 ± 47.76	0.049 ± 0.108

Fig. 2. Experiment environment. Participant wearing headphones and holding the DualSense controller.

Kruskal-Wallis test confirmed significant differences between conditions ($H = 33.80$, $p < 0.001$), and a Mann-Whitney U test indicated the proposed method had significantly higher correlation than the NOISE condition ($U = 122.0$, $p = 0.004$).

3.2 Qualitative Evaluation

We conducted a user study with 10 participants (8 male, 2 female, aged 19–25, mean age 20.5), all with normal hearing and vibrotactile sensitivity. Audio was presented through noise-cancelling headphones, and vibrotactile feedback was simultaneously delivered via a DualSense Wireless Controller[2](Figure 2). The stimuli consisted of 9 audio scenes selected from the 12 used in the quantitative evaluation and classified into three categories: **Impact Events** (ball bouncing, door slamming, footsteps), **Continuous Textures** (engine, rain, floor scraping), and **Complex Scenes** (conversation + gravel walking, screaming + gunshots, station ambience). Participants evaluated the experience on a 7-point Likert

[2] Sony Interactive Entertainment: https://www.playstation.com/accessories/dualsense-wireless-controller/

scale across five dimensions based on the HX model [4]: *Realism*, *Harmony*, *Involvement*, *Expressivity*, and *Distraction* (lower is better).

Table 2. Qualitative Evaluation Results (Overall Mean ± SD)

Condition	*Realism*	*Harmony*	*Involvement*	*Expressivity*	*Distraction*↓
Proposed	4.53±1.64	**4.67**±1.70	**4.43**±1.80	**4.29**±1.76	**2.61**±1.39
Lowpass	**4.54**±1.53	4.54±1.59	4.39±1.66	4.28±1.62	2.66±1.48
Envelope	4.09±1.68	4.58±1.61	4.19±1.73	4.12±1.76	3.37±2.05
Noise	3.58±1.73	3.50±1.66	3.30±1.85	3.23±1.81	2.87±1.60

The results (Table 2) indicate that the proposed method achieved the highest or equivalent scores on all positive evaluation items and recorded the lowest Distraction value. A Friedman test revealed significant differences between conditions for all evaluation items ($p < 0.05$ or better). Post-hoc comparisons using the Wilcoxon signed-rank test with Bonferroni correction showed significant differences between the Proposed and Noise conditions on all positive items ($p < 0.001$). While no significant differences were found between the Proposed method and the Lowpass condition on any items, a significant difference was observed between the Proposed method and the Envelope condition specifically for the Distraction item ($p < 0.001$).

Category-wise analysis revealed performance variations across stimulus types. For **Impact Events**, the proposed method achieved the highest Harmony (4.97) and Involvement (4.87) scores, demonstrating strong performance for transient stimuli. For **Continuous Textures**, the proposed method led in Realism (4.57) and Involvement (4.33), with the lowest Distraction among non-noise conditions (2.57). For **Complex Scenes**, the proposed method achieved the highest Realism (4.27) and Expressivity (4.07), suggesting effective handling of multi-component audio events.

4 Discussion

The quantitative and qualitative evaluation results confirm that the proposed dual-branch model can generate temporally synchronized audio and vibration from text prompts. The mean onset time difference of 54.45 ms falls below the perceptual synchrony threshold (55 ms), suggesting that parallel decoding from the shared latent space contributes to maintaining temporal synchrony. Qualitatively, the proposed method achieved the highest or equivalent scores on all positive evaluation items, with the lowest Distraction score (2.61) indicating natural and non-intrusive vibration generation.

Statistical tests showed no significant differences between the proposed method and rule-based methods on most items, indicating equivalent subjective quality under the experimental conditions (200 Hz LPF, 55 Hz envelope). However,

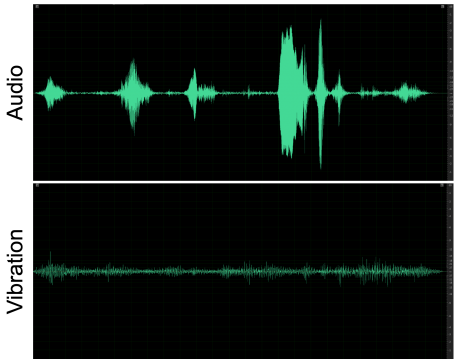

Fig. 3. Learned voice suppression. For a prompt "children laughing in a playground," the audio branch outputs voice components while the vibration branch suppresses them, demonstrating semantic filtering learned from Spleeter-preprocessed training data.

the proposed method significantly outperformed the Envelope condition on Distraction ($p < 0.001$), suggesting that learning-based transformation suppresses excessive vibration responses.

The category-wise analysis provides additional insights. The proposed method showed particular strengths in Impact Events (highest Harmony and Involvement), suggesting effective capture of temporal dynamics for transient stimuli. For Complex Scenes with voice components, the proposed method maintained competitive performance, likely due to the semantic filtering capability learned through Spleeter-preprocessed training data. Continuous Textures showed the smallest performance gaps between methods, consistent with the inherent difficulty of vibration representation for sustained sounds.

If equivalent quality can be achieved with rule-based methods, it might seem sufficient to simply apply rule-based transformation after audio generation. However, rule-based methods convert only physical characteristics (low-frequency components, envelope), whereas a learning-based VibrationDecoder can utilize semantic information in the latent space. Additionally, rule-based transformation propagates audio artifacts directly to vibration, while the dual-branch architecture allows haptic-specific optimization through DPO fine-tuning.

Compared with recent learning/knowledge-driven approaches—Scene2Hap [3] (LLM + physical modeling over VR scenes) and Sound2Hap [6] (audio-to-vibration learned from human ratings)—both assume a VR description or an audio track is already given and treat alignment as a downstream problem. Our method instead generates both modalities jointly from text via a shared diffusion latent, making synchrony a structural property; this framing keeps the approach valuable even with numerical scores on par with rule-based baselines, since the same backbone extends uniformly to other modalities and inherits future improvements of text-to-audio diffusion for free.

Figure 3 demonstrates an example of semantic filtering learned by the VibrationDecoder. When a prompt containing voice elements is input (children's playground, with laughter), the audio branch generates output that includes the voice component, while the vibration branch significantly suppresses the gain for voice frequencies. This behavior emerges from training on pseudo-vibration data that excludes voice components (via Spleeter preprocessing), indicating that the VibrationDecoder has learned to filter out content inappropriate for haptic feedback. Such semantic-aware suppression would be difficult to achieve with simple rule-based low-pass filtering, which cannot distinguish between voice and other mid-frequency content.

Key limitations include training on pseudo-vibration data rather than real tactile recordings; real-time performance constraints requiring acceleration via Consistency Models; and single-channel output limiting spatial vibration patterns. Because the pseudo-vibration targets follow a fixed LPF–envelope rule, the VibrationDecoder inherits its spectral/temporal biases—high-frequency microtextures above 200 Hz are unseen and sustained sounds are represented mainly through the 55 Hz carrier—which partly explains why the gap to rule-based baselines is smallest for **Continuous Textures** and largest for **Impact Events**. The DPO stage uses only 200 preference pairs as an initial investigation, and scaling to thousands of pairs should further tighten the Distraction gains. We also observed non-negligible inter-participant variability, especially for complex scenes, suggesting future evaluations adopt mixed-effects analysis. Future work will address these through real-environment data collection, larger preference datasets, and multi-actuator extensions.

5 Conclusion

This study proposed a dual-branch latent diffusion model that simultaneously generates sound effects and vibrotactile feedback from text prompts. The core contribution is the architectural design where audio and vibration are generated from a shared latent space, structurally guaranteeing temporal synchronization. Built upon Make-An-Audio 2, the system generates vibration spectrograms through a VibrationDecoder (Transformer-CNN architecture). Evaluation results showed that the proposed method achieved a mean onset time difference of 54.45 ms (below the 55 ms threshold) and subjective quality equivalent to rule-based methods. The significantly lower Distraction score suggests that learning-based transformation produces more natural vibrations. Future work includes training with real audio-tactile correspondence data, real-time generation using Consistency Models, and extension to spatial vibration patterns.

Acknowledgments. We thank the Information Somatics Laboratory, Research Center for Advanced Science and Technology (RCAST), The University of Tokyo, for their cooperation with the ethics review of this study. We also thank Kenji Yamahiro for his help with the operation of the user study.

References

1. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 33, pp. 6840–6851 (2020)
2. Huang, J., Ren, Y., Huang, R., Yang, D., Ye, Z., Zhang, C., Liu, J., Yin, X., Ma, Z., Zhao, Z.: Make-an-audio 2: Temporal-enhanced text-to-audio generation. arXiv preprint arXiv:2305.18474 (2023)
3. Jingu, A., AliAbbasi, E., Safaee, S., Strohmeier, P., Steimle, J.: Scene2hap: Generating scene-wide haptics for vr from scene context with multimodal llms. In: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems. pp. 1–21 (2026)
4. Kim, E., Schneider, O.: Defining haptic experience: Foundations for understanding, communicating, and evaluating hx. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–13. ACM (2020)
5. Kim, S., Lee, G., Lee, K.: Hapticgen: Generative text-to-vibration model for haptic feedback design. In: CHI Conference on Human Factors in Computing Systems (CHI '25). ACM (2025)
6. Li, Y., Seifi, H.: Sound2hap: Learning audio-to-vibrotactile haptic generation from human ratings. In: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems. pp. 1–19 (2026)
7. Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., Plumbley, M.D.: Audioldm: Text-to-audio generation with latent diffusion models. In: International Conference on Machine Learning (ICML). pp. 21450–21474. PMLR (2023)
8. Occelli, V., Spence, C., Zampini, M.: Audiotactile interactions in temporal perception. Psychonomic Bulletin & Review **18**(3), 429–454 (2011)
9. Petry, B., Huber, J., Nanayakkara, S.: Scaffolding the machine: Investigating and improving vibrotactile feedback for cross-modal audio-tactile displays. IEEE Transactions on Haptics **11**(2), 213–224 (2018)
10. Precision Microdrives: Audio-to-vibe: Understanding low-pass filters for haptic feedback. Application Bulletin 003 (2020)
11. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 36 (2023)
12. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (2022)
13. Slater, M., Wilbur, S.: A framework for immersive virtual environments five: Speculations on the role of presence in virtual environments. Presence: Teleoper. Virtual Environ. **6**(6), 603–616 (Dec 1997)
14. Stein, B.E., Stanford, T.R.: Multisensory integration: current issues from the perspective of the single neuron. Nature Reviews Neuroscience **9**(4), 255–266 (2008)
15. Ujitoko, Y., Ban, Y.: Tactgan: Vibrotactile design using generative adversarial network. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–7. ACM (2021)
16. Yamaguchi, K., Konyo, M., Tadokoro, S.: Sensory equivalence conversion of high-frequency vibrotactile signals using intensity segment modulation method for enhancing audiovisual experience. In: 2021 IEEE World Haptics Conference (WHC). pp. 674–679 (2021)