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

# Perceptual Compression of Multimodal Tactile Signals with an Attention-Enhanced Autoencoder and Cross-Modal Psychohaptic Loss Function

Wenxuan Wei, Xiao Xu, Lars Nockenberg, Daniel Rodriguez-Guevara, Eckehard Steinbach
School of Computation, Information and Technology, Department of Computer Engineering,
Chair of Media Technology and Munich Institute of Robotics and Machine Intelligence (MIRMI),
Technical University of Munich, Munich, Germany
{wenxuan.wei, xiao.xu, lars.nockenberg, daniel.rodriguez, eckehard.steinbach}@tum.de

***Abstract*—This paper presents MPTC-Net, an autoencoder-based perceptual codec for multimodal tactile signals, capable of jointly compressing data across multiple tactile dimensions. Previous studies, including the state-of-the-art vibrotactile codecs standardized in IEEE 1918.1.1 and MPEG-I Haptics Coding, have primarily focused on roughness-related information, rather than jointly encoding multiple tactile dimensions. To address this limitation, we developed a Multimodal Psychohaptic Model (MPM) that incorporates the impact of multimodal stimulation on perceptual thresholds. The MPM is integrated into the loss function during training to enhance perceptual performance. Furthermore, an attention module is employed to extract critical information across modalities, and both early fusion and late fusion strategies are explored for improved multimodal integration. Our experimental results show significant improvements with the proposed codec, particularly in vibrotactile perceptual metrics, demonstrating its effectiveness in managing the complexity of multimodal tactile feedback.**

***Index Terms*—Tactile codec, Perceptual coding, Multimodal fusion, Tactile Internet.**

## I. Introduction

The Tactile Internet aims to provide realistic touch experiences and immersive multi-sensory remote exploration of real or virtual environments [1]. Transmitting multimodal haptic data is crucial, as objects exhibit diverse physical properties, and the human sensory system processes multiple haptic stimuli concurrently to perceive and interact with the environment accurately [2]. For example, vibrotactile feedback enables the perception of fine textures, contact force feedback supports the recognition of shape and hardness, and frictional forces convey the degree of surface roughness. These capabilities are crucial for applications like teleoperation, VR, and E-commerce, ensuring immersive and realistic interactions.

This work was funded by the European Union's Horizon Europe research and innovation program (HORIZON-MSCA-2022-DN-01) under the Marie Skłodowska-Curie grant agreement No. 101073465 (TOAST). It also received support from the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy – EXC 2050/1 (Project ID 390696704) – through the Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) at Technische Universität Dresden. Additional support was provided in part by the Sino-German Mobility Programme (M-0421).

Recent multimodal research has investigated the interactions and integration of different sensory modalities. Some studies have explored cross-modal correlations between vision and touch [3], [4], while others have focused on multimodal fusion [5], [6]. Our work focuses on efficient compression and transmission of multiple tactile stimuli.

Tactile stimuli are associated with surface properties, comprising five key dimensions of tactile perception: macro and fine roughness, warmness/coldness, hardness/softness, and friction [7]. Various sensing and display technologies for multi-modal feedback have been explored, such as Texplorer2 [8], Proton Pack [9], a tool-mediated recording device [10] and multimodal haptic gloves [11]. However, the large amount of multimodal tactile data generated by these devices presents real-time transmission challenges. For example, vibrotactile signals for 36 points require a transmission rate of 1.6 Mbit/s [12], and adding other tactile modalities further increases the bitrate. Therefore, achieving efficient transmission while preserving perceptual transparency depends on effective compression of multimodal tactile signals. Multimodal codecs exploit the relationships between different modalities and human perception, making them a key research focus.

Among the various forms of tactile data, vibrotactile information has attracted the most attention, with standards already established by both IEEE [13] and MPEG [14]. Several vibrotactile codecs have been developed, which can be broadly categorized into two types: transform-based codecs and deep learning-based (DL) codecs.

Transform-based codecs compress vibrotactile signals by converting them into the frequency domain and applying frequency-dependent, perceptually optimized quantization to preserve relevant details. Two state-of-the-art transform-based approaches are PVC-SLP [15] and VC-PWQ [16]. PVC-SLP uses sparse linear prediction for perceptual coding, while VC-PWQ adopts a discrete wavelet transform and psychohaptic model for quantization. An extended version of VC-PWQ supports multiple interaction points [17].

Deep learning-based codecs are gaining attention for their ability to automatically extract meaningful features and support end-to-end optimization of the entire compression

pipeline. These advantages can help achieve higher compression ratios while maintaining reconstruction quality. Li et al. [18] proposed a CNN-based codec optimizing the rate distortion function. In [19], the recurrent network-based vibrotactile codec (RNVC) shows the benefits of low latency. Nockenberg et al. developed a perceptually trained ResNet-based autoencoder [20] and its rate-scalable version [21], which we refer to as DeepVib. The above methods are effective, but they only focus on the vibrotactile data and neglect other tactile modalities. One recent approach to multimodal compression employed a stacked autoencoder (SAE) combined with a gated recurrent unit (GRU) to encode normal and lateral forces [22]. However, the absence of a perceptual model may compromise perceptual quality, as such a model is crucial for prioritizing information most relevant to human perception.

In summary, most existing tactile codecs focus only on vibrotactile signals, lacking multimodal integration and perceptual consideration. To our knowledge, our work is the first to propose a perceptually trained multimodal codec for vibrotactile, normal, and tangential forces—crucial elements to capturing surface properties like roughness, hardness, and friction. Although force signals are typically considered kinaesthetic, they also convey rich tactile cues from the surface. Developing such a codec poses challenges in signal integration, efficient feature extraction, and perceptual modelling. To address these challenges, we explore early and late fusion strategies to enhance signal integration, implement attention mechanisms to improve feature extraction, and propose a multimodal psychohaptic model to optimize the perceptual quality of multimodal interaction. The main contributions of this paper are:

- Development of an enhanced attention-based ResNet multimodal codec to efficiently compress multimodal tactile data.
- Proposal of a multimodal psychohaptic model that accounts for cross-modal influences in the codec design.
- Comprehensive evaluation of different fusion configurations and training loss functions, validated through metrics-based performance assessment.

## II. Background: DL-Based Vibrotactle Codec

The deep learning codecs comprise four main components: encoder block, quantization, entropy coding and decoder block. As illustrated in Fig. 1: (i) The encoder block takes a vibrotactile signal $\boldsymbol{x}$ as input and outputs a latent representation $\boldsymbol{y}$; (ii) The quantization module reduces the bits needed to describe the latent representation $\boldsymbol{y}$, but introducing errors to the compressed output space $\hat{\boldsymbol{y}}$ ; (iii) The entropy coding uses statistical models to further compress the data and generate a bitstream; (iv) Lastly, the decoder block reconstructs the time-domain signal $\hat{\boldsymbol{x}}$, from the compressed $\hat{\boldsymbol{y}}$.

Our multimodal codec builds upon the ResNet-based vibrotactile codec structure introduced in [20]. As shown in Fig. 2, the encoder and decoder each consist of a series of residual blocks. The encoder and decoder have a flexible design to

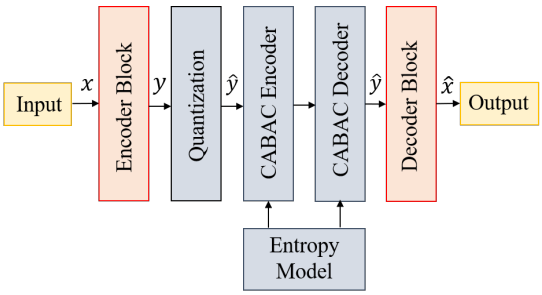
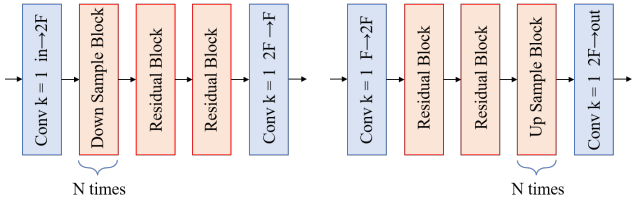



Fig. 1: DL-based tactile codec structure. Adapted from [20].

adjust the depth of the neural network and compression ratio. The size of the latent variables depends on the sampling factor $N$ and the number of features $F$, and it can be calculated as

$$y_{\text{size}} = F \times \frac{bl}{2^N} \tag{1}$$

where $bl$ represents the block length. Uniform quantization is applied after the encoder. Gaussian distribution is used to derive the entropy model. Context Adaptive Binary Arithmetic Coding (CABAC) is employed for entropy coding. This codec provides a flexible structure that allows adjustments to the model depth and compression ratio, making it adaptable to different use cases. However, it is limited to only supporting vibrotactile signals and relies on simple convolution blocks without advanced mechanisms such as attention. These limitations mark possibilities for improvement in our proposed multimodal codec.




(a) Encoder Block (b) Decoder Block

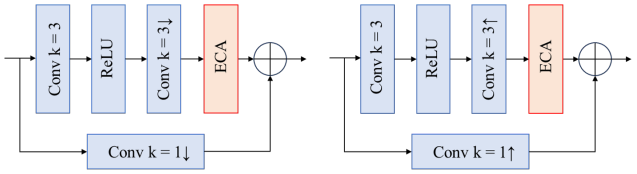
Fig. 2: Encoder and decoder blocks. The parameters are specified as "kernel size, input features → output features". Two additional residual blocks further extract high-level features after downsampling. Adapted from [20].

## III. The Multimodal Codec

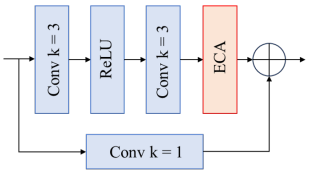
We propose key innovations to achieve a multimodal tactile codec and improve its performance: (1) an attention mechanism to capture complex data interdependencies, (2) exploration of early and late fusion techniques for modality integration, and (3) a multimodal psychohaptic model to develop a perceptual loss function.

### A. Attention Module

Fig. 3 shows the structure of the downsampling, upsampling and residual block. The difference is whether they include convolutional layers for upsampling or downsampling. Each downsampling block applies downsampling by a factor of 2,




(a) Downsampling Block (b) Upsampling Block




(c) Residual Block

Fig. 3: Residual Block Structures

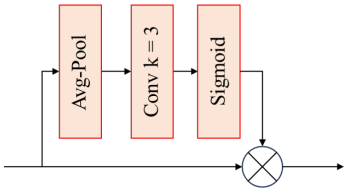


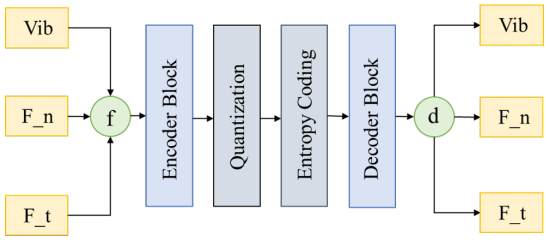

Fig. 4: ECA Module

while the upsampling block does the opposite. The residual blocks increase the network depth and further extract high-level features. We incorporate an Efficient Channel Attention (ECA) module [23] into the residual block by placing it before the residual connections, as shown in Fig. 3 and Fig. 4. By applying attention weights across different channels, the ECA module helps the network to focus on more informative features and ignore less relevant information. Moreover, the applied ECA is lightweight, enabling seamless integration into residual blocks with minimal impact on processing speed.

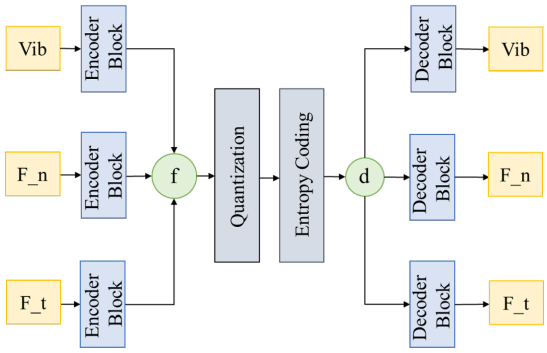
### *B. Multimodal fusion*

Multimodal fusion is a key challenge in designing efficient codec architectures, affecting reconstruction quality and compression efficiency. Depending on the fusion position within the neural network, the corresponding methods can be divided into early and late fusion. Early fusion merges data at the beginning of processing, allowing cross-correlations to be exploited and simplifying the model architecture. Late fusion processes each modality independently before combining the outputs. This approach offers advantages in scenarios where the quality of data varies across modalities or when significant differences exist between them. The purpose of comparing these two strategies is to determine which approach achieves more effective integration of features from different modalities, thereby enhancing overall compression performance.

Fig. 5 illustrates the two fusion strategies. The early fusion method (Fig. 5a) concatenates the three types of signals along the feature dimension and then feeds them into a shared encoder network. After quantization and entropy encoding, latent variables are processed together by a decoder block. Finally, the compressed signals are decoupled along the feature




(a) Early Fusion




(b) Late Fusion

Fig. 5: Two multimodal fusion strategies. Vib, F_n, and F_t correspond to vibrotactile, normal force, and tangential force. The letters 'f' and 'd' represent fusion and defusion.

dimension to reconstruct the three signals. On the other hand, the late fusion (Fig. 5b) encodes the three signals independently through separate encoder blocks, and their respective results are concatenated before the quantization. Similarly, the multimodal tactile information is decoupled and passed through individual decoder blocks for reconstruction.

Before feeding the force signals into the network, preprocessing is applied to capture dynamic features. Specifically, the mean value (DC component) is subtracted from each signal. The DC component is excluded from training and transmitted separately via header encoding, ensuring the model focuses on dynamic features while effectively integrating them with vibration signals.

### *C. Multimodal Psychohaptic Model*

The psychohaptic model describes how physical stimuli, such as force and vibration, are sensed and interpreted by the human. Noll et al. [16] introduced a vibrotactile psychohaptic model to adapt the quantizer, ensuring that distortions are introduced mainly in frequency ranges that are least perceptible. They analyzed the magnitude range of 280 vibrotactile signals collected from [24], identified the key threshold levels at specific frequencies to derive the absolute threshold function. The vibrotactile signals were obtained under a normal force of 1N. As illustrated in Fig. 6, the red curve represents the absolute threshold function:

$$t(f) = \left| \frac{62\text{dB}}{\left(\log_{10}\left(\frac{6}{11}\right)\right)^2} \cdot \left[\log_{10}\left(\frac{f}{550\text{Hz}} + \frac{6}{11}\right)\right]^2 \right| - 77\text{dB} \tag{2}$$

where $t(f)$ represents the absolute threshold as a function of the frequency $f$. Eq. (2) illustrates the relationship between frequency and the vibrotactile absolute threshold. However, the influence of other modalities is not considered. Since material texture perception involves contact and pressing, we need to extend the threshold function by incorporating the impact of normal force. As a preliminary exploration, our study specifically focuses on the effects of normal force, while the influence of tangential force on other modalities remains an open question.

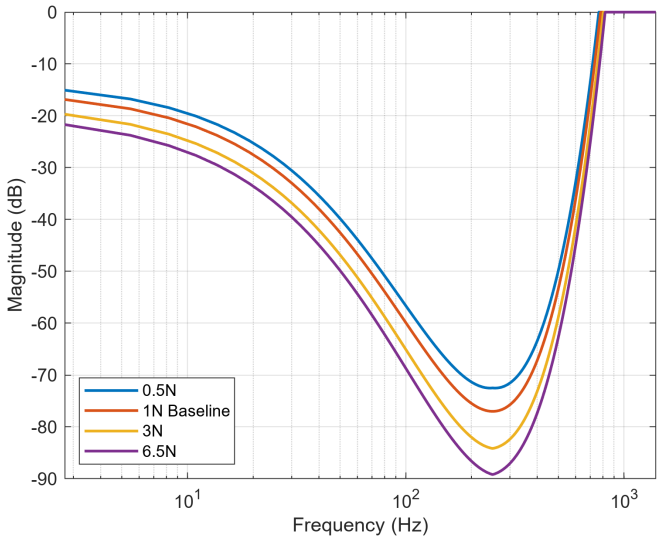



Fig. 6: Absolute threshold affected by normal force. The previous absolute threshold curve with a 1N normal force is from [16](red), while the other three curves represent absolute thresholds under different normal forces: 0.5 N (blue), 3 N (yellow), and 6.5 N (purple).

The stimuli in one modality influences the perceptual threshold of another modality. This is recognized as the "cross-modality threshold artifacts" [25], [26]. For example, a stronger normal force leads to a significant decreased vibration threshold in the normal direction, but the threshold of tangential vibration is independent from the normal force [26]. Oh et al. found that the impact of force on perception thresholds is more noticeable at a frequency of 250 Hz [25]. Meanwhile, the result from [25] shows that at a vibration frequency of 250 Hz, the effect of force on the perception thresholds can reach up to 10 dB at a vibration frequency of 250 Hz, while it is typically within 5 dB at 40 Hz. Using this information as regression points, we improved the function by adding a variable $F$ to represent the effect of normal force.

$$t(f,F) = \left| \frac{62\text{dB}}{\left(\log_{10}\left(\frac{6}{11}\right)\right)^2} \cdot \left[\log_{10}\left(\frac{f}{550\text{Hz}} + \frac{6}{11}\right)\right]^2 \right| - \frac{15 \cdot \log_{10}(F)}{1 + \left|\log_{10}\left(\frac{f}{255} + \frac{1}{51}\right)\right|} - 77\text{dB} \tag{3}$$

In Eq. (3), $F$ represents the normal force. The added part of the function reflects the influence of normal force on the vibrotactile threshold. Fig. 6 illustrates the new absolute threshold function for vibrotactile signals considering the cross-modality artifacts. The red curve is considered as the baseline as presented in [16] with a 1N-normal force. The remaining three curves correspond to different levels of normal force (0.5N, 3N and 6.5N). The curves indicate that pressing force has a stronger influence on the threshold in the mid-frequency range, i.e., 100-500 Hz. This ensures that our model aligns with the findings reported in [25]. In the next subsection, we use this psychophysical model to design the loss function, enabling adaptability to complex multimodal interactions, which the original model in [16] could not achieve.

### *D. Loss Function*

The loss function for deep learning-based codecs consists of two major components: entropy loss and distortion loss. The entropy loss encourages the model to produce more compact representations. Distortion loss focuses on the accuracy of the reconstructed data and is typically calculated by MSE. Perceptual loss is also considered a distortion loss, focusing on the perceptual quality of the reconstructed data. In this section, we introduce the psychohaptic model and the perceptual loss function derived from it.

Mask-to-Noise Ratio (MNR) is a concept introduced in the Psychohaptic Model (PM) introduced in [16] and can be used to measure how much reconstruction noise can be perceived. PM includes both the absolute perceptual threshold and masking effects. In [16], the two components are added by power additive combination to obtain the global mask. According to the results of [20], when the compression ratio is not very high, the masking effect has little impact on the signal quality. This means neglecting masking under such conditions does not lead to significant degradation. Hence, in this work, we simplified the method by using Eq. (3) from the previous section as the global threshold. Specifically, for each vibrotactile sample, the normal force $F$ in Eq. (3) is calculated based on the average of the samples over the same period. The vibrotactile signal is divided into multiple wavelet bands in the frequency domain. For each band, MNR can be calculated as follows:

$$MNR(b) = 10\log_{10}\frac{E_{\text{g}}(b)}{E_{\text{n}}(b)} \tag{4}$$

where $b$ represents the band index. $E_{\text{g}}(b)$ and $E_{\text{n}}(b)$ denote the global mask and the noise energy across the band $b$. The vibrotactile perceptual loss $L_{MNR}$ for vibrotactile signals is based on [20], and obtained from:

$$L_{MNR} = -\frac{1}{20}\sum_{b=1}^{B} MNR(b) - 22 \tag{5}$$

where $B$ is the total number of bands. Please note, that compared to [16], the MNR used in this paper uses the $t(f,F)$ from Eq. (3) and not the $t(f)$ shown in Eq. (2).

The Perceptual Mean Squared Error metric (P-MSE) [27] is used as the distortion loss for normal force and tangential

force. PMSE uses Weber-Fechner law to describe the relationship between psychophysical sensation and physical stimulus. It is defined as

$$L_{PMSE} = \frac{1}{N}\sum_{i=0}^{N-1}\left[S(i) - \hat{S}(i)\right]^2 = \frac{c^2}{N}\sum_{i=0}^{N-1}\left[\log\frac{x_i}{\hat{x}_i}\right]^2 \quad (6)$$

$S$ and $\hat{S}$ represent the original and distorted psychophysical sensations, while $x$ and $\hat{x}$ denote the corresponding force values in time domain. The constant $c$ is a scaling factor determined experimentally.

Based on the above calculations, we have the perceptual loss function for vibrotactile and force signals. Specifically, $L_{MNR}$ represents the perceptual distortion in vibrotactile signals while $L_{PMSE}$ serves as a specialized loss function for force signals. To evaluate the effectiveness of the proposed loss function used in the multimodal codec, we compared the two loss functions: Signal Loss ($L_S$) and Perceptual Loss ($L_P$). Signal loss uses MSE as distortion loss for all modalities. Perceptual loss aims to balance pixel-level accuracy with perceptual quality.

$$L_S = \lambda_1 MSE_{vib} + \lambda_2 MSE_{F_n} + \lambda_2 MSE_{F_t} + \lambda_3 R \quad (7)$$

$$L_P = \lambda_1 MSE_{vib} + (1-\lambda_1)L_{MNR} + \lambda_2 L_{PMSE\_F_n} + \lambda_2 L_{PMSE\_F_t} + \lambda_3 R \quad (8)$$

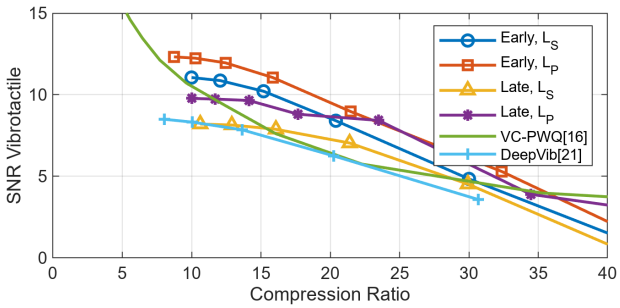
$\lambda_1$, $\lambda_2$ and $\lambda_3$ are the coefficients to balance between the terms.

## IV. Experiments

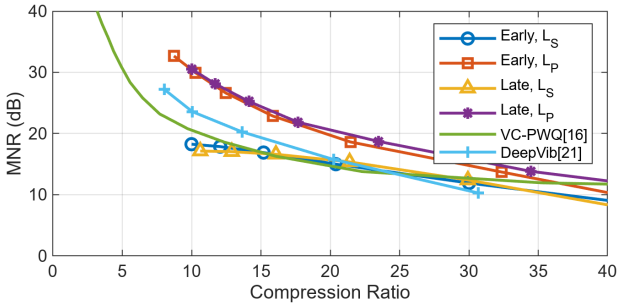
### A. Dataset and Parameters

We used the dataset by Culbertson et al. [10], which includes time-aligned vibration, normal force, and tangential force signals for 100 haptic textures, each recorded for 10 seconds at 10 kHz using a custom handheld device. To match the IEEE 1918.1.1 standard [13], signals were downsampled to 2800 Hz and split into ten 1-second segments, with 8 segments used for training and 2 for validation.

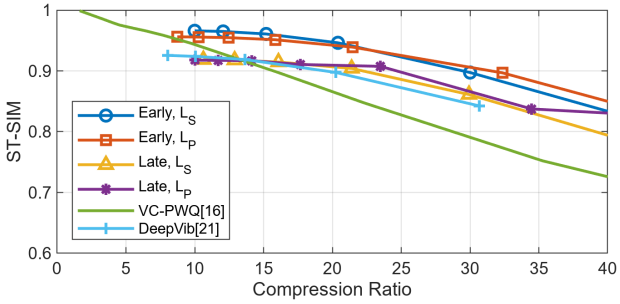
We used the Adam optimizer with an exponential scheduler and set the number of downsampling blocks $N$ to 3. For early fusion, the number of the features $F$ is set to 6, to balance compression ratio and signal quality. For late fusion, the feature numbers for vibration and force are 4 and 1, respectively. Since the vibrotactile signal is more complex, we assign 4 features. This configuration ensures that the overall complexity and compression ratio range are aligned to compare. In Eq. (7), $\lambda_1$ and $\lambda_2$ are both set to 1. In Eq. (8), the values of $\lambda_1$ and $\lambda_2$ are set to 0.9 and 1, respectively, to balance the influence of MSE and perceptual component. $\lambda_3$ is set to $5 \times 10^{-8}$ because the entropy loss has a much larger scale than the distortion loss. The network parameters were initialized with He initialization. The other parameters are: batch_size = 128, iterations = 300, learning rate = $1 \times 10^{-3}$.



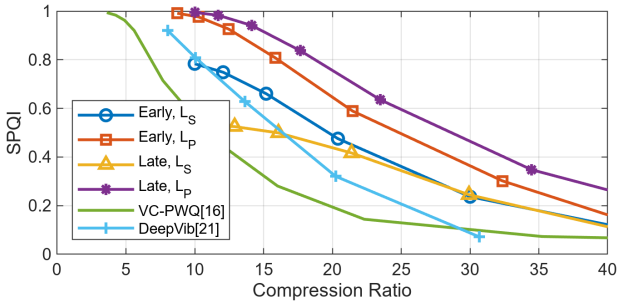

(a) SNR




(b) MNR based on Eq.(3)




(c) ST-SIM




(d) SPQI

Fig. 7: Vibrotactile Quality Metrics

### B. Evaluation

This part presents a comparative analysis of training results under different configurations. Specifically, we evaluate the performance of four setups, combining early and late fusion techniques with two different loss functions ($Ls$ from Eq. (7) and $Lp$ from Eq. (8)). The evaluated metrics are Signal-to-Noise Ratio (SNR), MNR, Spectral Temporal SIMilarity (ST-SIM), and Spectral Perceptual Quality Index (SPQI). VC-PWQ codec [16] and the rate-scalable deep learning-based codec [21], which we refer to as DeepVib, are adopted as the

baselines for the vibrotactile signal. To the best of the authors' knowledge, there are no previous results in the literature for evaluating the tactile codec performance for normal and tangential forces. As a result, the comparative experiments for these modalities lack a baseline for reference.

*1) Vibrotactile Evaluation:* Fig. 7 shows metric results vs. compression ratio for vibrotactile signals. As seen in Fig. 7a, MPTC-Net outperforms VC-PWQ [16] and DeepVib [21] within the 15–30 compression ratio range. Moreover, using perceptual loss $L_p$ yields better SNR than signal loss $L_s$.

MNR, plotted in Fig. 7b, is an evaluation metric that indicates the perceptibility of reconstruction noise within a specific frequency band. The two curves that use perceptual loss (red and purple) demonstrate the best performance, compared to the baselines and the results using signal loss. Although DeepVib also used perceptual loss, our methods achieve greater improvements. The results obtained using signal loss perform similarly to the VC-PWQ method.

Spectral Temporal SIMilarity (ST-SIM) is a Vibrotactile quality assessment method that includes perceptual spectral and temporal similarity measures [28]. In Fig. 7c, MPTC-Net outperforms baselines across most compression ratios, particularly at high compression levels. Early fusion results are slightly better than late fusion.

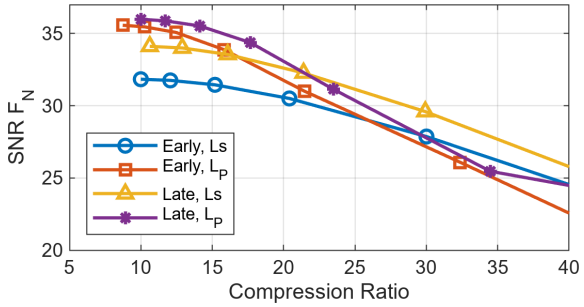
Spectral Perceptual Quality Index (SPQI) computes a similarity score based on a computed perceptually weighted error measure [29]. It is an effective metric for assessing subjective quality, where higher values indicate better performance. The results shown in Fig. 7d indicate that the perceptual loss outperforms both signal loss and VC-PWQ. Deep learning methods show a clear improvement over VC-PWQ, and our approach outperforms DeepVib at high compression ratios. Among the evaluated methods, late fusion achieves the best overall performance.

*2) Force Evaluation:* We evaluate both the normal and tangential forces using the SNR metric. As shown in Fig. 8, the SNR results vary from 20 to 35 and reflect a reliable quality for compression ratios between 10 to 40. For high compression ratios, the quality of normal force is slightly better than that of tangential force. In the dataset, the normal forces are more stable, while the tangential forces have some randomness and greater complexity, which may cause slightly lower reconstruction quality.

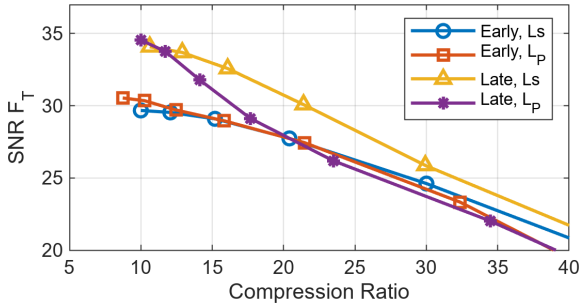
In conclusion, all four setups show comparable performance, except early fusion with signal loss, which performs worse. As shown in Fig. 8b, late fusion achieves better reconstruction of tangential force, likely due to its superior ability to reduce cross-modal interference.

### C. Discussion

Regarding vibrotactile signals, our codec with perceptual loss performs better in perceptual metrics, especially in SPQI, which aligns well with subjective experimental results [29]. In Fig. 7a, unlike VC-PWQ, the curve shows no upward trend at low compression rates due to inherent distortion from the deep learning-based method. Allocating more bits in entropy coding offers limited performance gains, so training multiple models to span a broader range of compression ratios is meaningful.




(a) Normal Force




(b) Tangential Force

Fig. 8: Force Evaluation

Perceptual metrics allow us to evaluate vibrotactile signal quality without time-consuming user studies. However, since SNR may not capture perceived force signal quality, future work should include subjective tests to assess the multimodal codec more comprehensively.

In summary, perceptual loss demonstrates significant improvements over signal loss and the baselines, particularly when the compressed signals are evaluated using perceptual metrics. Meanwhile, both early fusion and late fusion exhibit superior performance on specific metrics, with the optimal choice depending on practical considerations such as transmission efficiency and model complexity.

## V. Conclusion

In this paper, we proposed a ResNet-based multimodal tactile codec to efficiently compress vibrotactile, normal force, and tangential force. First, we considered the impact of multiple modalities on the perception threshold of the vibrotactile signal and proposed a multimodal psychohaptic model to design a perceptual loss function. We further introduced early and late fusion strategies for multimodal signals and perceptually trained deep learning models. Evaluation using different quality metrics shows that the multimodal perceptual tactile codec outperforms the baselines VC-PWQ and deepVib, especially in vibrotactile perceptual quality. According to the experiments, the perceptual loss and late fusion strategies improved performance for the perceptual quality metrics. Future research could further validate the effectiveness through user experiments, extending the modalities to include temperature, and applying it to VR and teleoperation systems.

## REFERENCES


[1] E. Steinbach, M. Strese, M. Eid, X. Liu, A. Bhardwaj, Q. Liu, M. Al-Ja'afreh, T. Mahmoodi, R. Hassen, A. El Saddik, and O. Holland, "Haptic codecs for the tactile internet," *Proceedings of the IEEE*, vol. 107, no. 2, pp. 447–470, 2019.

[2] D. Wang, K. Ohnishi, and W. Xu, "Multimodal haptic display for virtual reality: A survey," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 1, pp. 610–623, 2020.

[3] H. Lu, K. Yang, W. Zhuang, and M. Chen, "Vision-assisted cross-modal haptic data compression for immersive communication," in *2024 IEEE Wireless Communications and Networking Conference (WCNC)*, 2024, pp. 1–6.

[4] Q. Tong, W. Wei, C. Liu, X. Zhou, Y. Zhang, and D. Wang, "Cross-modal transmission with active packet loss and restoration for tactile internet," *IEEE Communications Magazine*, vol. 62, no. 8, pp. 70–76, 2024.

[5] Q. Wang, Y. Bai, and H. Song, "Middle fusion and multi-stage, multi-form prompts for robust rgb-t tracking," *Neurocomputing*, vol. 596, p. 127959, 2024.

[6] F. Fooladgar and S. Kasaei, "Multi-modal attention-based fusion model for semantic segmentation of rgb-depth images," *arXiv preprint arXiv:1912.11691*, 2019.

[7] S. Okamoto, H. Nagano, and Y. Yamada, "Psychophysical dimensions of tactile perception of textures," *IEEE Transactions on Haptics*, vol. 6, no. 1, pp. 81–93, 2013.

[8] M. Strese, L. Brudermueller, J. Kirsch, and E. Steinbach, "Haptic material analysis and classification inspired by human exploratory procedures," *IEEE Transactions on Haptics*, vol. 13, no. 2, pp. 404–424, 2020.

[9] A. Burka and K. J. Kuchenbecker, "How much haptic surface data is enough?" in *2017 AAAI Spring Symposium Series*, 2017.

[10] H. Culbertson, J. J. López 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)*, 2014, pp. 319–325.

[11] Q. Tong, W. Wei, Y. Guo, T. Jin, Z. Wang, H. Zhang, Y. Zhang, and D. Wang, "Distant handshakes: Conveying social intentions through multi-modal soft haptic gloves," *IEEE Transactions on Affective Computing*, pp. 1–15, 2024.

[12] A. Noll, B. Gulecyuz, A. Hofmann, and E. Steinbach, "A Rate-scalable Perceptual Wavelet-based Vibrotactile Codec," in *2020 IEEE Haptics Symposium (HAPTICS)*. IEEE, 2020, pp. 854–859. [Online]. Available: https://ieeexplore.ieee.org/document/9086333/

[13] "Ieee standard for haptic codecs for the tactile internet," *IEEE Std 1918.1.1-2024*, pp. 1–127, 2024.

[14] P. Guillotel, Y. Muthusamy, Q. Galvane, E. Vezzoli, L. Nockenberg, I. Sodagar, H. Da Costa, A. Hulsken, G. Lecuyer, M. P. Da Silva, F. Bouffard, H. Culbertson, S. Kollannur, and D. Gueorguiev, "Adding touch to immersive media: An overview of the mpeg haptics coding standard," *IEEE Transactions on Haptics*, pp. 1–15, 2025.

[15] R. Hassen, B. Gülecyüz, and E. Steinbach, "Pvc-slp: Perceptual vibrotactile-signal compression based-on sparse linear prediction," *IEEE Transactions on Multimedia*, vol. 23, pp. 4455–4468, 2021.

[16] A. Noll, L. Nockenberg, B. Gulecyuz, and E. Steinbach, "VC-PWQ: Vibrotactile Signal Compression based on Perceptual Wavelet Quantization," in *2021 IEEE World Haptics Conference (WHC)*. IEEE, 2021, pp. 427–432.

[17] L. Nockenberg, A. Noll, S. Panëels, A. B. Dhiab, C. Hudin, and E. Steinbach, "Mvibcode: Multi-channel vibrotactile codec using hierarchical perceptual clustering," *IEEE Transactions on Haptics*, vol. 16, no. 4, pp. 646–651, 2023.

[18] Z. Li, R. Hassen, and Z. Wang, "Autoencoder for Vibrotactile Signal Compression," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 4290–4294.

[19] T. Zhao, Y. Fang, K. Wang, Q. Liu, and Y. Niu, "High efficiency vibrotactile codec based on gate recurrent network," *IEEE Transactions on Multimedia*, vol. 25, pp. 5043–5052, 2023.

[20] L. Nockenberg and E. Steinbach, "Vibrotactile signal compression using perceptually trained autoencoders," in *International Conference on Human Haptic Sensing and Touch Enabled Computer Applications*. Springer, 2024, pp. 264–277.

[21] L. Nockenberg, W. Wei, M. Navai, and E. Steinbach, "Deep learning-based perceptual vibrotactile codec with rate scalability," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[22] G. Liu, X. Li, C. Wang, and S. Lv, "Online compression and reconstruction for communication of force-tactile signals," *IEEE Communications Letters*, vol. 27, no. 3, pp. 981–985, 2023.

[23] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 534–11 542.

[24] J. Kirsch, A. Noll, M. Strese, Q. Liu, and E. Steinbach, "A low-cost acquisition, display, and evaluation setup for tactile codec development," in *2018 IEEE International Symposium on Haptic, Audio and Visual Environments and Games (HAVE)*, 2018, pp. 1–6.

[25] S. Oh and S. Choi, "Effects of Contact Force and Vibration Frequency on Vibrotactile Sensitivity During Active Touch," *IEEE Transactions on Haptics*, vol. 12, no. 4, pp. 645–651, Oct. 2019.

[26] Y. D. Pra, S. Papetti, H. Järveläinen, M. Bianchi, and F. Fontana, "Effects of vibration direction and pressing force on finger vibrotactile perception and force control," *IEEE Transactions on Haptics*, vol. 16, no. 1, pp. 23–32, 2023.

[27] R. Chaudhari, E. Steinbach, and S. Hirche, "Towards an objective quality evaluation framework for haptic data reduction," in *2011 IEEE World Haptics Conference*, 2011, pp. 539–544.

[28] R. Hassen and E. Steinbach, "Subjective evaluation of the spectral temporal similarity (st-sim) measure for vibrotactile quality assessment," *IEEE Transactions on Haptics*, vol. 13, no. 1, pp. 25–31, 2020.

[29] A. Noll, M. Hofbauer, E. Muschter, S.-C. Li, and E. Steinbach, "Automated quality assessment for compressed vibrotactile signals using multi-method assessment fusion," in *2022 IEEE Haptics Symposium (HAPTICS)*, 2022, pp. 1–6.