HapticLLaMA: A Multimodal Sensory Language Model for Haptic Captioning

Anonymous ACL submission

Abstract

Haptic captioning is the task of generating natural language descriptions from haptic signals, such as vibrations, for use in virtual reality and rehabilitation applications. While previous multimodal research has focused primarily on vision and audio, haptic feedback for the sense of touch remain underexplored. To address this gap, we formalize the haptic captioning task and propose HapticLLaMA, a multimodal sensory language model that interprets vibration signals into descriptions in a given sensory, emotional, or associative category. We investigate two types of haptic tokenizers, a frequency-based tokenizer and an EnCodec-based tokenizer, that convert haptic signals into sequences of discrete units, enabling their integration with the LLaMA model. HapticLLaMA is trained in two stages: (1) supervised fine-tuning using the LLaMA architecture with LoRA-based adaptation, and (2) fine-tuning via reinforcement learning from human feedback (RLHF). We assess HapticLLaMA's captioning performance using both automated n-gram metrics and human evaluation. HapticLLaMA demonstrates strong capability in interpreting haptic vibration signals, achieving a METEOR score of 59.98 and a BLEU-4 score of 32.06, respectively. Furthermore, over 64% of the generated captions received human ratings above 3.5 on a 7-point scale, with RLHF yielding a 13% improvement in the overall rating distribution, indicating stronger alignment with human haptic perception. These findings highlight the potential of large language models to process and adapt to sensory data.

1 Introduction

Humans perceive their environment through five primary senses: vision, hearing, touch (haptics), taste, and smell. Integrating these sensory modalities can enhance AI systems' ability to interpret human perception and behavior by providing a richer and more nuanced understanding of user context and sensory experiences. Haptic signals convey diverse information through tactile sensations, including physical attributes such as surface texture (Culbertson et al., 2014), emotional states like urgency or pleasantness (Yoo et al., 2015), and recognizable real-world cues such as heartbeats or buzzing of a bee (Seifi et al., 2015; Seifi and MacLean, 2017). These tactile interactions have broad applications in areas including user interactions in virtual reality (VR), physical rehabilitation, blind user navigation, and gaming (Choi et al., 2021; Seim et al., 2022; Katzschmann et al., 2018; Yun et al., 2023a; MacLean et al., 2017).

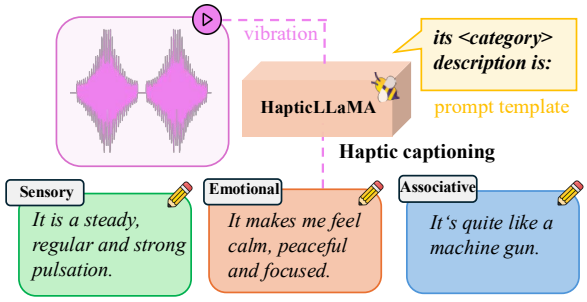


Figure 1: HapticLLaMA can generate sensory, emotional, and associative captions for an input vibration haptic signal with a tailored prompt template.

Multimodal Large Language Models (MLLMs) have significantly advanced captioning across various modalities such as images (Vinyals et al., 2015; Mokady et al., 2021), video (Iashin and Rahtu, 2020; Wu et al., 2023), and audio (Zhang et al., 2022; Liu et al., 2022). *Haptic captioning* involves generating textual descriptions (captions) of haptic signals, capturing sensory, emotional, and associative qualities (see Figure 1). This process enhances the interpretability of physical signals by describing human perception of vibration patterns, similar to how vision-language models capture human perception of images. Additionally, haptic captioning offers a novel opportunity to develop datasets and

benchmarks that probe the limits of AI capabilities, challenging models to accurately interpret and articulate complex physical sensations through natural language.

Unlike image and audio captioning, haptic captioning remains an underexplored domain. As a nascent field, haptic-language understanding presents unique challenges but also significant potential for advancing multimodal AI. The task involves two key challenges: (1) the lack of established tokenization methods for representing vibration haptic signals in a format suitable for language models; and (2) the absence of sensory multimodal models capable of processing vibration haptics. To date, little work has investigated the ability of large language models (LLMs) to interpret haptic signals.

To support the haptic captioning task, we introduce **HapticLLaMA**, a multimodal haptic language model trained with two distinct haptic tokenizers that integrates haptic vibration signals and textual descriptions within a single framework. Raw haptic signals are continuous time-series data, incompatible with the discrete token-based input required by LLMs. To address this, we investigate two types of haptic tokenizers: a frequency-based tokenizer and an EnCodec-based tokenizer (Défossez et al., 2023), both designed to convert vibrations into interpretable token sequences suitable for LLMs. HapticLLaMA is trained in two stages: (1) supervised fine-tuning using a LoRA-adapted (Hu et al., 2022) LLaMA architecture (Touvron et al., 2023), and (2) reinforcement learning from human feedback (RLHF; Ouyang et al., 2022) applied using generated haptic captions rated by humans.

Specifically, we contribute: (1) HapticLLaMA, a haptic language model capable of generating sensory, emotional, and associative captions for vibration signals, (2) two haptic tokenizers that convert raw vibration inputs into sequences of discrete tokens, (3) the VibRate dataset, containing 16,896 user-rated <vibration, caption, rating> samples, used to further incorporate human perception into the model via RLHF, and (4) extensive experiments using automated n-gram metrics and human evaluation demonstrating that HapticLLaMA exhibits strong perceptual capabilities in describing vibration signals, achieving a METEOR score of 59.98 and a BLEU-4 score of 32.06, and receiving human ratings above 3.5 (on a 7-point scale) for over 64% of the generated captions.

2 Related Work

2.1 Haptic Modality and Touch Datasets

Early research on haptic language understanding (Obrist et al., 2013; Seifi et al., 2015; Seifi and MacLean, 2017; Dalsgaard et al., 2022) emerged within the human-computer interaction (HCI) domain, primarily through qualitative studies. Recent work proposed large-scale touch-language datasets for robotic sensing. Datasets such as TVL (Fu et al., 2024), TLV (Cheng et al., 2024b), Touch100k (Cheng et al., 2024a), and ObjectFolder (Gao et al., 2023) use tactile images captured by deformable RGB-like sensors (Yuan et al., 2017). These datasets convey object properties like shape, size, and texture. Similarly, MimicTouch (Yu et al.) collects tactile images from human demonstrations of object interaction to train robots in object manipulation. While useful for training robots to perceive physical objects, these datasets lack programmable haptic feedback like vibrations from phones and VR controllers, which are common in user-facing applications. For vibration haptics, Hu et al. (2024) developed a pipeline for mapping emotional tags to haptic signal features, but it was only tested on 32 signals with 12 descriptions each. Hu et al. (2025) introduced HapticCap, a dataset of 92,070 vibration samples paired with sensory, emotional, and associative descriptions, which we use as a basis to train HapticLLaMA.

2.2 Multimodal Captioning

Multimodal models for image (Vinyals et al., 2015; Mokady et al., 2021), video (Iashin and Rahtu, 2020; Wu et al., 2023), and audio captioning (Zhang et al., 2022; Liu et al., 2022) have rapidly advanced in the last decade. Recent advances in open-source LLMs such as LLaMA (Touvron et al., 2023) have accelerated the development of multimodal models. In the haptic–language domain, prior work has centered on assistive applications or haptic signal generation rather than producing rich text descriptions of vibration inputs. Wang et al. (2023) combines vibration notifications with auditory captions in an assistive system; HapticVLM (Khan et al., 2025) generates audio and vibration from images to emulate surface textures; Lu et al. (2022) generates texture vibrations based on user preference ratings; and Scene2Hap (Jingu et al., 2025) produces vibrations from images of a virtual reality scene. In contrast, we present HapticLLaMA, a sensory language model that generates

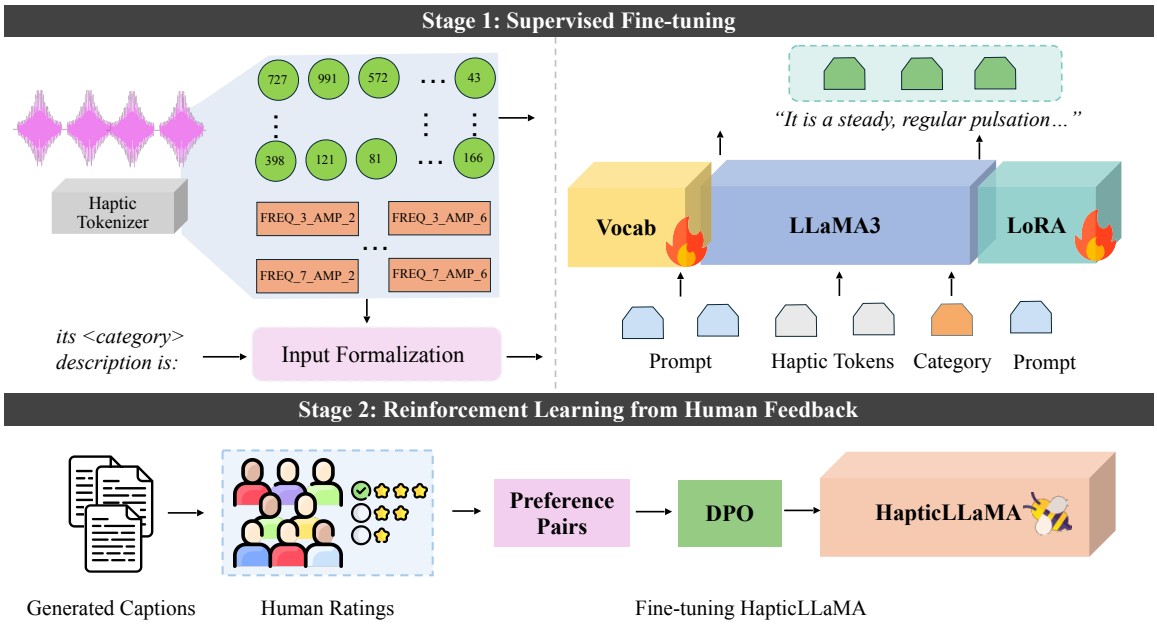


Figure 2: Overview of the two-stage process for constructing HapticLLaMA: (1) Supervised fine-tuning, where haptic tokens from two alternative tokenizers are formatted into a multimodal prompt to fine-tune a LLaMA model using LoRA; and (2) Reinforcement learning from human feedback (RLHF), where user ratings for generated captions from Stage 1 are converted into preference pairs for DPO fine-tuning, yielding the final HapticLLaMA.

textual captions directly from vibration inputs, advancing haptic signal understanding in the absence of suitable vibration foundation-model baselines.

3 HapticLLaMA

3.1 Task Definition

Similar to image (Vinyals et al., 2015; Mokady et al., 2021) and audio captioning (Zhang et al., 2022; Liu et al., 2022), haptic captioning involves generating textual descriptions (i.e., captions) from haptic signals. Given a vibration signal S and a target category $c \in \{\text{sensory}, \text{emotional}, \text{associative}\}$, where sensory refers to physical attributes (e.g., intensity of tapping), emotional denotes affective impressions (e.g., the mood of a scene), and associative indicates real-world familiar experiences (e.g., buzzing of a bee, a heartbeat), the goal is to generate a caption corresponding to the specified category of haptic experience.

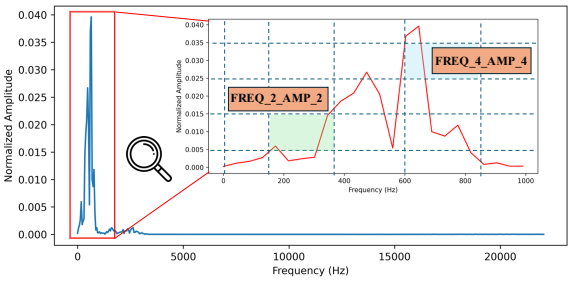
3.2 Overall Architecture

As shown in Figure 2, HapticLLaMA is built on the LLaMA architecture and consists of a haptic tokenizer, a LLaMA model enhanced with LoRA, and a human feedback module trained using reinforcement learning. We process the haptic signals offline, converting them into sequences of discrete tokens using two methods: a frequency-based tokenizer using spectral frequency information and an EnCodec-based pretrained neural audio codec (Défossez et al., 2023). The haptic tokens and target category are formatted into a multimodal prompt and input to a LLaMA model fine-tuned with LoRA. Following supervised fine-tuning, human ratings of the generated captions are collected to further refine the model via Direct Preference Optimization (DPO; Rafailov et al., 2023).

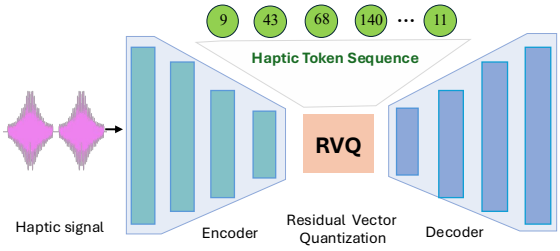
3.3 Input Formalization

3.3.1 Haptic Tokenizer

Frequency-based Tokenizer: This tokenizer (Figure 3(a)) is motivated by the importance of spectral frequency information in characterizing vibrations (Jones and Sarter, 2008; Bensmaïa et al., 2005). Unlike the time domain, where signals are represented by amplitude values over time, the frequency domain represents a signal as a sum of sinusoids at varying frequencies. The frequency-based tokenizer begins by converting the time-domain signal into the frequency domain via the Fast Fourier Transform (FFT), and subsequently discretizes the resulting frequency components through binning into variable-width intervals. Our proposed tokenizer divides the frequency range into logarithmically spaced bins that correspond to just-noticeable differences in human frequency perception (Choi



(a) Frequency-based tokenizer.



(b) EnCodec-based tokenizer.

Figure 3: Two haptic tokenizers to encode vibration signals into input representations suitable for LLaMA.

and Kuchenbecker, 2013; Israr et al., 2006). Similarly, the amplitude range is segmented into normalized levels. The tokenizer then assigns a unique token (e.g., FREQ_3_AMP_2) to each frequency-amplitude pair, encoding the signal's spectral content into a form interpretable by LLMs.

EnCodec-based Tokenizer: EnCodec[1] is a neural audio codec that compresses audio using deep learning (Défossez et al., 2023). It consists of three main components: (1) an encoder that transforms raw audio into a lower-dimensional latent representation, (2) a quantizer that discretizes the latent features via residual vector quantization (Zeghidour et al., 2021), and (3) a decoder that reconstructs the waveform from the quantized codes. Since vibrations have perceptual and rhythmic similarities to audio signals (Bernard et al., 2022; Pätzold et al., 2023; Yun et al., 2023b; Degraen et al., 2021), we adopt the discrete codes produced by the quantizer as haptic tokens for HapticLLaMA (Figure 3(b)).

3.3.2 Input Format

After tokenization, the haptic tokens are added to the vocabulary of LLaMA tokenizer and their embeddings are updated in the learning process. Given an input tuple $\langle s, c \rangle$, where s represents the signal token sequence generated by the haptic tokenizer and c denotes the category given from $\{$sensory, emotional, associative$\}$. For the input of HapticLLaMA, we concatenate haptic, category, textual prompt, and special tokens (e.g., <EOS>) as a multimodal prompt I. We use the prompt I="`haptic signal: <haptic tokens>, its <category> description is: <`*caption*`>.`". During training, we append the human-written caption as reference. The auxiliary text tokens (e.g., *"its"*, *"description"*, and *"is"*) can be interpreted as prompts that guide the model's output.

3.4 Multimodal HapticLLaMA Model

Algorithm 1 outlines the two-stage training procedure of HapticLLaMA, consisting of (1) supervised fine-tuning with LoRA adaptation and (2) subsequent fine-tuning based on human feedback on generated captions, as detailed below.

3.4.1 Stage 1: Supervised Fine-Tuning

We adopt LLaMA3[2] as the backbone of HapticLLaMA. To align the haptic and text modalities, we incorporate haptic tokens into the LLaMA tokenizer's vocabulary by using randomly initialized special tokens in the LLaMA vocabulary. Their embeddings are updated during training, enabling the language model to effectively interpret and utilize the haptic tokens. For efficient fine-tuning, we employ Low-Rank Adaptation (LoRA; Hu et al., 2022), which inserts trainable low-rank matrices ΔW_* into the model weights.

$$\begin{aligned} W_*^{\text{lora}} &= W_* + \Delta W_* \\ &= W_* + B_* A_*, * \in \{Q, V\} \end{aligned} \tag{1}$$

where W_*^{lora} denotes the LoRA-adapted weight matrix, and W_* refers to the original pretrained weight matrix (e.g., the query Q or value V projection weights in a transformer). The subscripts $* \in \{Q, V\}$ indicate the Query and Value projection layers, respectively, and only the matrices A_* and B_* are trainable parameters.

3.4.2 Stage 2: Fine-Tuning via RLHF

Using the trained model from Stage 1, we generate captions for each vibration signal in **VibRate** dataset (See Section 4 for details) across sensory, emotional, and associative categories and collect user ratings for the captions on 1-7 Likert scale. The rated captions are then paired into preferred and rejected captions for fine-tuning using DPO[3].

[1] https://huggingface.co/facebook/encodec_24khz

[2] https://huggingface.co/meta-llama/Llama-3.2-3B

[3] https://huggingface.co/docs/trl/dpo_trainer

Algorithm 1: HapticLLaMA Training

Require: HapticCap Train set $T = (S, D)$, VibRate signal set S', describe category c, prompt p, LLaMA, haptic tokenizer.

Ensure: Every signal is associated with a corresponding description.

Stage 1: Supervised Fine-Tuning

1: Load pretrained LLaMA.
2: Tokenize each haptic signal into discrete haptic tokens $S = \{s_1, \ldots, s_n\}$.
3: Include haptic tokens and update LLaMA's vocabulary.
4: Formalize input based on S, p, and c into $[p_1, \cdots, s_i, \cdots, p_i, \cdots,$ <EOS>$]$.
5: Supervised fine-tune HapticLLaMA with formalized input.
6: **return** initial HapticLLaMA $\mathcal{H}$.

Stage 2: Fine-Tuning via RLHF

7: Infer the caption D' of $s \in S'$ using trained $\mathcal{H}$.
8: Construct the VibRate dataset by incorporating human ratings R.
9: Construct caption preference pairs set $\hat{T}$ based on VibRate.
10: Fine-tune HapticLLaMA $\mathcal{H}$ with $\hat{T}$.
11: **return** The final HapticLLaMA🪄.

Preference Pairs: To create preference pairs for DPO, we match captions rated above the midpoint of the scale with those rated below it. Rated captions are split into high-rated (positive) and low-rated (negative) groups using a 3.5 threshold. Each high-rated caption is set as the preferred choice and paired with all low-rated captions as the non-preferred (rejected) ones. The HapticLLaMA model is then trained to assign higher likelihood to the preferred response over the rejected one, aligning generation with human preferences.

4 VibRate Dataset Construction

VibRate is a diverse, manually curated dataset of 16,896 <vibration, caption, rating> tuples (see Appendix F for details on recruitment and payment). It includes 704 vibrations constructed from four diverse sources: (a) 174 vibration signals are created by varying signal parameters (e.g., frequency); (b) 180 vibrations derived from sound effect libraries by mimicking timing or applying low-pass filtering (Ternes and MacLean, 2008; Degraen et al., 2021; Yun et al., 2023b); (c) 176 vibrations generated by HapticGen (Sung et al., 2025); and (d) 174 custom-made vibrations created manually through signal transformations such as time reversal, repetition, and mixing (Schneider and MacLean, 2016; MacLean et al., 2017). For each signal, we generate four captions in each category: two using frequency-based $\mathcal{H}$ and two from EnCodec-based $\mathcal{H}$. Then, we collect ratings from 44 human evaluators for the generated captions. Annotators are instructed to assign a final rating on a scale of 1 (poor) to 7 (excellent) based on two criteria: (1) the clarity and semantic accuracy of the caption, and (2) the alignment between the haptic vibration experience and caption. A higher rating indicates better quality and closer alignment with human perception. In total, 44 users participated, each evaluating captions for 32 different vibrations, yielding 16,896 <vibration, caption, human rating> samples.

5 Experiments

Datasets: We use two datasets to develop and evaluate HapticLLaMA: **HapticCap** (Hu et al., 2025) and **VibRate** on two stages respectively. **HapticCap** is the largest fully human-annotated haptic-captioned dataset, containing 92,070 haptic-text pairs, with 8-10 user-written captions per vibration, describing sensory, emotional, and associative attributes. The split details of HapticCap and VibRate datasets are given in Appendix B. We evaluate HapticLLaMA from both Stage 1 and Stage 2 on the test set (see Appendix C for implementation details).

Baselines: We set the following baselines: (1) **Random**: For each vibration in the test set, we randomly select one caption from the candidate set as the caption for that signal, (2) **Signal-agnostic**: Since LLMs can generate fluent descriptions grounded in their pretraining and guided through prompting, we examine the capability of LLaMA3.2-3B and GPT-4.5 [4] in generating captions without receiving any signal in the input (see Appendix A for details), (3) **Without LoRA Finetuning**: Fine-tuning is disabled by removing LoRA from the Frequency-based and EnCodec-based HapticLLaMA models, while keeping haptic token training. (4) **Without RLHF**: Model performance with and without DPO fine-tuning from human feedback is reported. Appendix E details the lack of comparable baselines in prior work.

[4] https://platform.openai.com/docs/models/gpt-4.5-preview

Models	BLEU-1	BLEU-4	ROUGE-L	METEOR
Random	5.77	2.59	10.62	11.07
LLaMA (signal-agnostic)	9.39	2.71	20.67	23.84
GPT-4.5 (signal-agnostic)	11.53	4.28	26.09	28.20
Frequency tokens + LLaMA	28.46	7.48	29.49	35.24
EnCodec tokens + LLaMA	30.07	8.61	29.50	35.26
Frequency tokens + LLaMA + LoRA	40.12±5.1	23.16±4.3	42.95±3.4	50.16±2.1
EnCodec tokens + LLaMA + LoRA	41.54±3.2	24.28±2.1	44.71±2.4	54.03±2.3
Frequency HapticLLaMA: Frequency tokens + LLaMA + LoRA + RLHF	49.43±1.4	30.86±1.5	48.74±2.1	58.95±2.1
EnCodec HapticLLaMA: EnCodec tokens + LLaMA + LoRA + RLHF	51.36±1.3	32.06±1.2	49.56±1.6	59.98±1.8

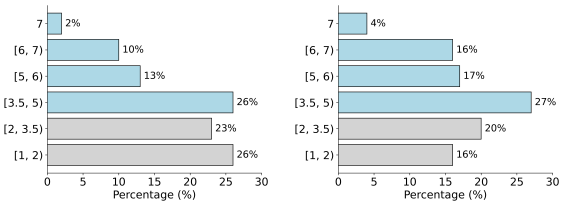
Table 1: Results on automated n-gram metrics. Baselines include Random, LLaMA, and GPT-4.5. Ablation results show the impact of each component and the performance of HapticLLaMA with Frequency and EnCodec tokens. Green and yellow denote the best and second-best performances, respectively. ± denotes standard deviation.

Evaluation Metrics: Following prior work on captioning (Drossos et al., 2020; Cui et al., 2018), we evaluate haptic captions using BLEU-1 (Papineni et al., 2002), BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) metrics. Due to the inherent ambiguity of haptic perception, each vibration in HapticCap has 8–10 ground-truth reference captions. We evaluate each prediction using the reference captions and report the mean and standard deviation of the metrics to reflect performance and variability. Since these text-based metrics cannot fully capture haptic semantics, we complement our evaluation with a human study. See Appendix D for correlation between automated metrics and human ratings.

5.1 Performance on Automated Metrics

Table 1 presents the performance of HapticLLaMA, the baselines, and the ablation results with different haptic tokenizers, LoRA adapter, and RLHF.

The random baseline exhibits consistently poor performance across all evaluation metrics. In contrast, the signal-agnostic LLaMA and GPT-4.5 models achieve marginal improvements, attributable to their advanced language generation capabilities despite the absence of haptic signals. Ablation results demonstrate the contribution of each component in HapticLLaMA. Incorporating haptic tokens into the training vocabulary leads to consistent improvements across all evaluation metrics. LoRA fine-tuning with haptic tokenizers shows strong absolute gains between +11.47 (BLEU-1) to +18.77 (METEOR) points in all metrics. Applying RLHF via DPO further boosts HapticLLaMA's performance across all metrics, including an absolute gain of +7.78 on BLEU-4 score. EnCodec tokens consistently outperform Frequency tokens by a small margin, with EnCodec HapticLLaMA (last row) achieving the highest scores across all metrics. This result may be because the Frequency-based tokenizer emphasizes the spectral domain while neglecting temporal and rhythmic features of vibration, whereas the EnCodec-based tokenizer captures richer representations of temporal changes based on the raw vibration signal and creates efficient tokens for signal compression and reconstruction. Additionally, the length of haptic token sequences and vocabulary size produced by the EnCodec tokenizer are significantly higher than that of the Frequency-based tokenizer (see Section 5.5), suggesting that EnCodec captures more fine-grained variations in the vibration signals. These results demonstrate that both frequency and EnCodec tokens can capture input signals, and the application of DPO further enhances caption generation quality by refining the alignment between model output and human preferences.



(a) After Stage 1. (b) After Stage 2.

Figure 4: Distribution of human ratings after Stage 1 and 2 for captions generated by EnCodec HapticLLaMA. Higher ratings are in blue and lower ones are in gray.

5.2 Human Evaluation

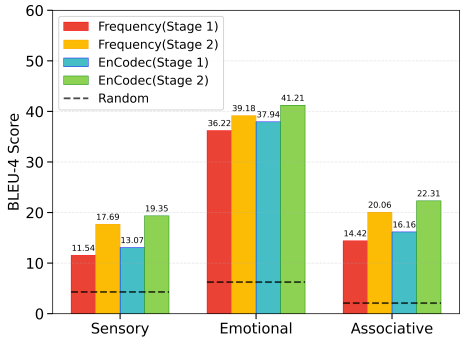
Figure 4 presents the human evaluation results of descriptions generated by HapticLLaMA in Stage 1 and Stage 2, respectively. We use the human ratings from VibRate as the evaluation results af-

ter Stage 1. After completing both training stages, we randomly select 50 vibration signals from the HapticCap test set and generate two haptic captions per description category using the final HapticLLaMA model for each of the EnCodec-based and Frequency-based model variants, resulting in 1,500 haptic captions. This sample size provides sufficient statistical power (>90%) to detect a medium effect (Cohen's d = 0.5) in a paired-samples t-test comparing ratings of the same signals across two models. Five evaluators are instructed to rate the captions on a 1–7 scale, using the same human rating setup as in Stage 1. We evaluate inter-rater reliability. The quadratic-weighted Cohen's kappa (Cohen, 1968; Landis and Koch, 1977) is 0.66 (substantial agreement), and the ICC(3,1) (Shrout and Fleiss, 1979) is 0.72 (high consistency), indicating that the human ratings are reliable and provide a sound basis for evaluation.

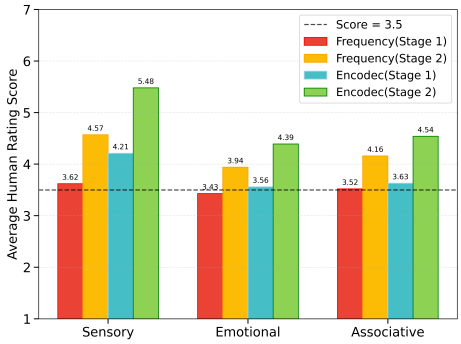
The results show a noticeable improvement in the quality of generated captions after Stage 2, with over 64% of ratings above 3.5. There is a visible shift in the distribution toward higher rating intervals, indicating that more captions are perceived as relevant and accurate after Stage 2. For instance, the proportion of higher ratings increases across all intervals after Stage 2, with about 13% increase in ratings above 3.5 and a particularly notable rise of about 6% and 4% in the range [6, 7) and [5, 6), respectively. In contrast, the share of lower ratings, such as [1, 2) and [2, 3.5), decreases significantly by around 10% and 3%, respectively. These results indicate that the preference alignment introduced in Stage 2 notably improves the model's ability to generate descriptions that are semantically clearer and perceptually aligned with human judgments.

5.3 Category-Specific Performance

Figure 5 compares HapticLLaMA variants across sensory, emotional, and associative categories using (a) BLEU-4 scores (with additional METEOR results in Appendix D) and (b) average human ratings. Both illustrate that the application of DPO consistently improves performance in all categories, yielding gains in BLEU-4 and average human rating, respectively, in line with the observed improvements in overall performance. As measured by BLEU-4, emotional captions yield the highest performance among the three categories across both HapticLLaMA variants, with the EnCodec-based model achieving the leading score of 41.2 in BLEU-4. This outcome may be



(a) BLEU-4 scores.



(b) Average human ratings.

Figure 5: EnCodec-based HapticLLaMA's performance across sensory, emotional, and associative categories.

attributed to the fact that emotional captions tend to show higher agreement among multiple human annotators (Hu et al., 2025), thereby facilitating more accurate and coherent generation.

According to human evaluations, sensory captions receive the highest ratings across all HapticLLaMA variants and stages. The discrepancy between automatic metrics and human judgments may stem from the limitations of automated methods (e.g., BLEU), which focus on surface-level textual similarity, whereas human raters are more attuned to the qualitative aspects of haptic experience. Similar discrepancy between automated metrics and human judgments are reported in image captioning literature (Elliott and Keller, 2014; Kilickaya et al., 2016).

5.4 Case Study

Figure 6 illustrates a comparison between reference ground-truth (R) and generated (G) captions for two haptic signals from the test set.

The left signal presents a continuous vibration featuring two prominent vibration peaks in the middle. Compared with the reference captions, the

Sensory	R: This sensation is a low intensity continuous vibration. G: It is continuous variable weak paced beat.	R: It feels like short and strong hit. G: The sensation was energetic with mild thumps in between.
Emotional	R: The sensation makes me feel a little anxious. G: This sensation made me feel nervous.	R: It feels energetic. G: It made me feel energetic and evoked excitement.
Associative	R: The sensation reminds me of a motorboat engine. G: This reminds me of damp shoes noise.	R: It reminded me of a ringing phone. G: It reminded me of an alarm clock.

Figure 6: Case study showing two example vibrations with captions. "R" (in green) denotes reference ground-truth captions from HapticCap, while "G" (in purple) denotes captions generated by EnCodec-based HapticLLaMA.

generated ones effectively capture continuity, weak intensity, and rhythm of the vibration in the sensory category. For the emotion dimension, the generated caption also successfully reflect emotional responses (e.g., "nervous" ↔ "anxious"). Regarding the associative category, HapticLLaMA broadens the range of associations from "motorboat engine" to "damp shoes", both capturing dull, heavy sensations and demonstrating a broader and more diverse associative space. The second signal displays features that are clearly distinct from those of the first. The vibration is an intermittent sequence of pulses with pauses at regular intervals. In the sensory caption, HapticLLaMA captures the pulsing characteristic ("short hit" ↔ "mild thumps in between") and overall energy ("strong" ↔ "energetic"). For the emotional aspect, HapticLLaMA aligns with the overall sentiment and enhances it (e.g., adding "evoked excitement"). In terms of associative categories, "ringing phone" and "alarm clock" sensations are logically similar, with both involving sequences of strong pulses in mobile phones.

5.5 Haptic Tokenizer Analysis

Table 2 presents the summary statistics of haptic tokens for Frequency and EnCodec tokenizers. EnCodec tokenizer has a much larger haptic token vocabulary (1,024) compared to the Frequency tokenizer (278), suggesting EnCodec can represent haptic signals with greater granularity, consistent with the improved performance observed in the EnCodec HapticLLaMA. The Frequency tokenizer produces shorter sequences for a vibration signal, with 47.5 tokens on average, indicating a more compact representation. In contrast, EnCodec outputs a fixed-length sequence of 1,379 tokens for all signals, which may introduce redundancy but ensures uniform input sizes for downstream models. In sum, the Frequency tokenizer offers higher compression efficiency, making it suitable for lightweight or real-time applications with a slightly lower performance, while EnCodec provides more detailed and consistent representations for tasks requiring richer signal understanding. The higher performance with EnCodec tokenizer is consistent with (Perianez-Pascual et al., 2025) and (Mousavi et al., 2025), extending audio-centric evidence on the strength of discrete tokenization (e.g., (Mousavi et al., 2025)) to haptic captioning.

Tokenizer	Item	Count
Frequency	Haptic token vocabulary size	278
	Average haptic signal length	47.5
	Max haptic signal length	52
	Min haptic signal length	12
EnCodec	Haptic token vocabulary size	1,024
	Average haptic signal length	1,379
	Max haptic signal length	1,379
	Min haptic signal length	1,379

Table 2: Summary statistics of haptic tokenizers.

Conclusion

We propose HapticLLaMA, a multimodal sensory language model trained with two distinct haptic tokenizers. Built on the LLaMA architecture, HapticLLaMA is trained in two stages: (1) supervised fine-tuning with LoRA, and (2) fine-tuning through RLHF. Our evaluation, combining automated generation metrics (e.g., BLEU and METEOR) and human assessments, shows that HapticLLaMA can effectively perceive vibration haptic signals and provide substantial improvements on the haptic captioning task. Our work advances haptic-language understanding by enabling large language models to interpret physical sensory data.

Limitations

BLEU, ROUGE, and METEOR metrics are suboptimal for evaluating haptic captioning quality, as they primarily emphasize textual fluency and lexical overlap, while failing to account for the semantic alignment between generated captions and the underlying haptic signals. HapticLLaMA focuses on vibration signals as the most accessible and diverse form of haptics, but cannot interpret other forms of haptics, such as force feedback or temperature signals. Finally, HapticLLaMA is trained on user descriptions and ratings from participants in a North American context. Because user background and culture can influence haptic descriptions, future work should systematically examine cultural and linguistic factors in vibration signal descriptions and extend the model to diverse cultures and languages.

Ethics Statement

While the idea of HapticLLaMA interpreting haptic vibration signals is appealing and holds potential for applications in human-computer interaction (HCI) and robotics, its current performance remains limited. The model achieves only around 32.06% on BLEU-4, an average human rating of 4.95 on a 7-point scale, and about 64% of captions rated above 3.5, indicating that its performance remains insufficient for real-world deployment.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Sliman Bensmaïa, Mark Hollins, and Jeffrey Yau. 2005. Vibrotactile intensity and frequency information in the pacinian system: A psychophysical model. *Perception & psychophysics*, 67(5):828–841.

Corentin Bernard, Jocelyn Monnoyer, Michaël Wiertlewski, and Sølvi Ystad. 2022. Rhythm perception is shared between audio and haptics. *Scientific Reports*, 12(1):4188.

Ning Cheng, Changhao Guan, Jing Gao, Weihao Wang, You Li, Fandong Meng, Jie Zhou, Bin Fang, Jinan Xu, and Wenjuan Han. 2024a. Touch100k: A large-scale touch-language-vision dataset for touch-centric multimodal representation. *arXiv preprint arXiv:2406.03813*.

Ning Cheng, You Li, Jing Gao, Bin Fang, Jinan Xu, and Wenjuan Han. 2024b. Towards comprehensive multimodal perception: Introducing the touch-language-vision dataset. *arXiv preprint arXiv:2403.09813*.

Inrak Choi, Yiwei Zhao, Eric J. Gonzalez, and Sean Follmer. 2021. Augmenting Perceived Softness of Haptic Proxy Objects Through Transient Vibration and Visuo-Haptic Illusion in Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics*, 27(12):4387–4400.

Seungmoon Choi and Katherine J. Kuchenbecker. 2013. Vibrotactile display: Perception, technology, and applications. *Proceedings of the IEEE*, 101(9):2093–2104.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. 2018. Learning to evaluate image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5804–5812.

Heather Culbertson, Juan José López Delgado, and Katherine J Kuchenbecker. 2014. One hundred data-driven haptic texture models and open-source methods for rendering on 3d objects. In *2014 IEEE haptics symposium (HAPTICS)*, pages 319–325. IEEE.

Tor-Salve Dalsgaard, Joanna Bergström, Marianna Obrist, and Kasper Hornbæk. 2022. A user-derived mapping for mid-air haptic experiences. *International Journal of Human-Computer Studies*, 168:102920.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2023. High fidelity neural audio compression. *Trans. Mach. Learn. Res.*, 2023.

Donald Degraen, Bruno Fruchard, Frederik Smolders, Emmanouil Potetsianakis, Seref Güngör, Antonio Krüger, and Jürgen Steimle. 2021. Weirding haptics: In-situ prototyping of vibrotactile feedback in virtual reality through vocalization. In *The 34th Annual ACM symposium on user interface software and technology*, pages 936–953.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE.

Desmond Elliott and Frank Keller. 2014. Comparing automatic evaluation measures for image description. In *52nd Annual Meeting of the Association for Computational Linguistics*, pages 452–457. Association for Computational Linguistics.

Letian Fu, Gaurav Datta, Huang Huang, William Chung-Ho Panitch, Jaimyn Drake, Joseph Ortiz, Mustafa Mukadam, Mike Lambeta, Roberto Calandra, and

Ken Goldberg. 2024. A touch, vision, and language dataset for multimodal alignment. *arXiv preprint arXiv:2402.13232*.

Ruohan Gao, Yiming Dou, Hao Li, Tanmay Agarwal, Jeannette Bohg, Yunzhu Li, Li Fei-Fei, and Jiajun Wu. 2023. The objectfolder benchmark: Multisensory learning with neural and real objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17276–17286.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Guimin Hu, Daniel Hershcovich, and Hasti Seifi. 2025. Hapticcap: A multimodal dataset and task for understanding user experience of vibration haptic signals.

Guimin Hu, Zirui Zhao, Lukas Heilmann, Yasemin Vardar, and Hasti Seifi. 2024. Grounding emotional descriptions to electrovibration haptic signals. *arXiv preprint arXiv:2411.02118*.

Vladimir Iashin and Esa Rahtu. 2020. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 958–959.

Ali Israr, Hong Z Tan, and Charlotte M Reed. 2006. Frequency and amplitude discrimination along the kinesthetic-cutaneous continuum in the presence of masking stimuli. *The Journal of the Acoustical society of America*, 120(5):2789–2800.

Arata Jingu, Easa AliAbbasi, Paul Strohmeier, and Jürgen Steimle. 2025. Scene2hap: Combining llms and physical modeling for automatically generating vibrotactile signals for full vr scenes. *arXiv preprint arXiv:2504.19611*.

Lynette A Jones and Nadine B Sarter. 2008. Tactile displays: Guidance for their design and application. *Human factors*, 50(1):90–111.

Robert K Katzschmann, Brandon Araki, and Daniela Rus. 2018. Safe local navigation for visually impaired users with a time-of-flight and haptic feedback device. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(3):583–593.

Muhammad Haris Khan, Miguel Altamirano Cabrera, Dmitrii Iarchuk, Yara Mahmoud, Daria Trinitatova, Issatay Tokmurziyev, and Dzmitry Tsetserukou. 2025. Hapticvlm: Vlm-driven texture recognition aimed at intelligent haptic interaction. *arXiv preprint arXiv:2505.02569*.

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2016. Re-evaluating automatic metrics for image captioning. *arXiv preprint arXiv:1612.07600*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Xubo Liu, Qiushi Huang, Xinhao Mei, Haohe Liu, Qiuqiang Kong, Jianyuan Sun, Shengchen Li, Tom Ko, Yu Zhang, Lilian H Tang, et al. 2022. Visually-aware audio captioning with adaptive audio-visual attention. *arXiv preprint arXiv:2210.16428*.

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):508–520.

Karon E MacLean, Oliver S Schneider, and Hasti Seifi. 2017. Multisensory Haptic Interactions: Understanding the Sense and Designing for It. In *The Handbook of Multimodal-Multisensor Interfaces: Foundations, User Modeling, and Common Modality Combinations*, volume 1, pages 97–142.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Pooneh Mousavi, Gallil Maimon, Adel Moumen, Darius Petermann, Jiatong Shi, Haibin Wu, Haici Yang, Anastasia Kuznetsova, Artem Ploujnikov, Ricard Marxer, et al. 2025. Discrete audio tokens: More than a survey! *arXiv preprint arXiv:2506.10274*.

Marianna Obrist, Sue Ann Seah, and Sriram Subramanian. 2013. Talking About Tactile Experiences. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 1659–1668.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Bastian Pätzold, Andre Rochow, Michael Schreiber, Raphael Memmesheimer, Christian Lenz, Max Schwarz, and Sven Behnke. 2023. Audio-based roughness sensing and tactile feedback for haptic perception in telepresence. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1387–1392. IEEE.

Jorge Perianez-Pascual, Juan D Gutiérrez, Laura Escobar-Encinas, Álvaro Rubio-Largo, and Roberto Rodriguez-Echeverria. 2025. Beyond spectrograms: Rethinking audio classification from encodec's latent space. *Algorithms*, 18(2):108.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Oliver S Schneider and Karon E MacLean. 2016. Studying design process and example use with macaron, a web-based vibrotactile effect editor. In *IEEE Haptics Symposium (Haptics)*, pages 52–58. IEEE.

Hasti Seifi and Karon E MacLean. 2017. Exploiting haptic facets: Users' sensemaking schemas as a path to design and personalization of experience. *International Journal of Human-Computer Studies*, 107:38–61.

Hasti Seifi, Kailun Zhang, and Karon E MacLean. 2015. Vibviz: Organizing, visualizing and navigating vibration libraries. In *2015 IEEE World Haptics Conference (WHC)*, pages 254–259. IEEE.

Caitlyn E. Seim, Brandon Ritter, Thad E. Starner, Kara Flavin, Maarten G. Lansberg, and Allison M. Okamura. 2022. Design of a Wearable Vibrotactile Stimulation Device for Individuals With Upper-Limb Hemiparesis and Spasticity. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 30:1277–1287. Conference Name: IEEE Transactions on Neural Systems and Rehabilitation Engineering.

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420.

Youjin Sung, Kevin John, Sang Ho Yoon, and Hasti Seifi. 2025. Hapticgen: Generative text-to-vibration model for streamlining haptic design. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–24.

David Ternes and Karon E MacLean. 2008. Designing large sets of haptic icons with rhythm. In *Haptics: Perception, Devices and Scenarios: 6th International Conference, EuroHaptics 2008 Madrid, Spain, June 10-13, 2008 Proceedings 6*, pages 199–208. Springer.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Yiwen Wang, Ziming Li, Pratheep Kumar Chelladurai, Wendy Dannels, Tae Oh, and Roshan L Peiris. 2023. Haptic-captioning: using audio-haptic interfaces to enhance speaker indication in real-time captions for deaf and hard-of-hearing viewers. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. 2023. Cap4video: What can auxiliary captions do for text-video retrieval? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10704–10713.

Yongjae Yoo, Taekbeom Yoo, Jihyun Kong, and Seungmoon Choi. 2015. Emotional responses of tactile icons: Effects of amplitude, frequency, duration, and envelope. In *2015 IEEE World Haptics Conference (WHC)*, pages 235–240. IEEE.

Kelin Yu, Yunhai Han, Qixian Wang, Vaibhav Saxena, Danfei Xu, and Ye Zhao. Mimictouch: Leveraging multi-modal human tactile demonstrations for contact-rich manipulation. In *8th Annual Conference on Robot Learning*.

Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. 2017. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762.

Gyeore Yun, Minjae Mun, Jungeun Lee, Dong-Geun Kim, Hong Z Tan, and Seungmoon Choi. 2023a. Generating Real-Time, Selective, and Multimodal Haptic Effects from Sound for Gaming Experience Enhancement. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, CHI '23, pages 1–17, New York, NY, USA. Association for Computing Machinery.

Gyeore Yun, Minjae Mun, Jungeun Lee, Dong-Geun Kim, Hong Z Tan, and Seungmoon Choi. 2023b. Generating real-time, selective, and multimodal haptic effects from sound for gaming experience enhancement. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.

Yiming Zhang, Hong Yu, Ruoyi Du, Zhanyu Ma, and Yuan Dong. 2022. Caption feature space regularization for audio captioning. *arXiv preprint arXiv:2204.08409*.

A Prompt for Signal-agnostic Experiments

In the absence of haptic signal input (signal-agnostic baseline), LLaMA is trained and tested purely on prompts *"its <category> description is: <caption>"* to evaluate its signal-agnostic capability. We provide the caption during the training and remove it from the prompt during inference. The prompts for GPT-4.5 are designed to guide the model in generating relevant haptic descriptions for sensory, emotional, and associative aspects (see Table 3), following the data collection guidelines in prior work (Hu et al., 2025). We provide several caption demonstrations in the prompt to guide the generation of GPT-4.5, as shown in Table 4.

	Role	Prompt
	system	You are experiencing a tactile sensation. Describe what you feel in one sentence.
Sensory	user	How would you describe the sensation?
Emotional	user	How does this sensation make you feel, can you attach any emotion to it?
Associative	user	Can you associate any action or object with this sensation?

Table 3: The prompt templates for GPT-4.5.

	Role	Response
Sensory	assistant	sensory: It is a steady, regular and strong pulsation
Emotional	assistant	emotional: It makes me feel calm, peaceful and focused.
Associative	assistant	associative: It's quite like a machine gun.

Table 4: The demonstrations for GPT-4.5.

B Details of Data Split

Table 5 presents the data split of the HapticCap dataset, along with the category distributions across sensory, emotional, and associative captions.

Category	Train	Valid	Test
Sensory	24,641	2,677	3,372
Emotional	24,641	2,677	3,372
Associative	24,641	2,677	3,372
All	73,923	8,031	10,116

Table 5: The details of HapticCap.

C Implementation Details

All experiments are conducted on NVIDIA RTX A100 and RTX H100 GPUs. Due to limited computational resources, we adopt LLaMA3.2-3B as the backbone of HapticLLaMA, and integrate Low-Rank Adaptation (LoRA) into all query and value projection layers of the Transformer architecture to enable parameter-efficient fine-tuning. In Stage 1, the batch size is set to 4, with an overall learning rate of 3e-4. The model is trained with a generative loss. In Stage 2, we use the default settings of the DPO trainer to optimize HapticLLaMA. The model is trained to align with human preferences through a simple classification objective, thereby improving caption quality.

D Additional Experimental Results

Figure 7 presents the METEOR scores of EnCodec HapticLLaMA across the three categories. The trends across the three categories are consistent with those observed in BLEU (see Figure 5(a)).

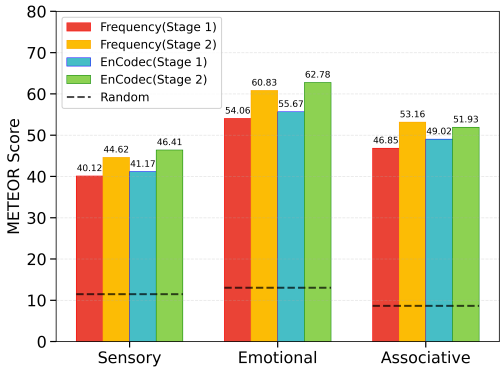


Figure 7: Performance of EnCodec-based HapticLLaMA across sensory, emotional, and associative categories, as measured by METEOR scores.

We calculate correlations between automatic metrics (e.g., BLEU) and human ratings to shed light on model performance, as shown in Tables 6 and Tables 7. Pearson and Kendall correlation results indicate that METEOR shows the strongest alignment with human ratings across sensory, emotional, and associative dimensions. The correlations remain at a moderate level ($r \approx$0.3–0.5), indicating that although lexical overlap metrics capture certain aspects of haptic descriptions, they may be limited in fully reflecting perceptual semantics.

Metric	Sensory	Emotional	Associative
BLEU-1	0.31	0.36	0.38
BLEU-4	0.26	0.27	0.29
ROUGE-L	0.35	0.39	0.42
METEOR	0.39	0.42	0.46

Table 6: Pearson correlation scores among BLEU-1, BLEU-4, ROUGE-L and METEOR.

Metrics	Sensory	Emotional	Associative
BLEU-1	0.26	0.33	0.39
BLEU-4	0.23	0.24	0.26
ROUGE-L	0.33	0.36	0.40
METEOR	0.35	0.38	0.43

Table 7: Kendall correlation scores among BLEU-1, BLEU-4, ROUGE-L and METEOR.

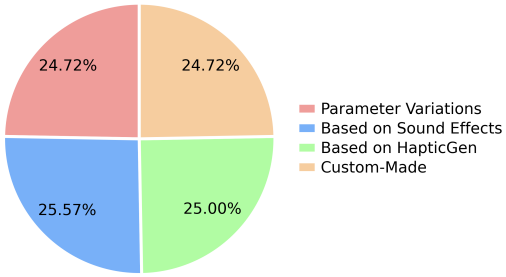


Figure 8: Haptic Source Distribution on VibRate.

E Lack of Comparable Datasets or Foundation Models

Cross-evaluation with other haptic datasets: Existing datasets, such as TVL (Fu et al., 2024), Touch100k (Cheng et al., 2024a), TLV (Cheng et al., 2024b), and ObjectFolder (Gao et al., 2023) focus on capturing visual–tactile associations such as shape and texture of an object, recorded through an RGB-based touch sensor. These datasets include tactile images of objects and do not include any vibration signals for comparison to our work. These resources primarily aim to bridge the gap between vision and touch, while HapticCap specializes in vibration haptic signals, tactile feedback perceived by humans through handheld devices (e.g., mobile phones, VR controllers). Thus, it is not possible to evaluate HapticLLaMA on these prior touch sensor datasets.

Comparison with other foundation models: The proposed HapticLLaMA is the first and only sensory model designed specifically for the vibration haptic captioning task, with no other available foundation models suitable for direct comparison. HapticVLM (Khan et al., 2025) is not a comparable model, as it is a vision-based system focused on material classification and temperature estimation based on images of a visual scene and object surfaces. Also, it generates sounds and vibrotactile feedback rather than generating user descriptions of haptic signals.

F Data Collection and Compensation

To collect the VibRate dataset, we curated vibration signals from different sources (see Section 4, Figure 8). Since real-world vibration recordings are often limited to surface textures and interactions, these vibration signals are manually created by human designers and collected from various sources to ensure signal richness and diversity. We recruited 44 university students as participants via advertisements and snowball sampling. Each participant felt 32 vibration signals on a Meta Quest 3 controller and rated their captions in a one-hour session. Following standard haptics practice, participants wore headphones and earplugs to block noise, held the Quest controller in their non-dominant hand, felt the vibration signals in random order, and rated twelve captions for each signal (3 caption categories$\times$2 captions$\times$2 HapticLLaMA variants). At the start of the ratings, participants were instructed to give low ratings when a sentence's meaning was unclear or did not match the vibration. To further calibrate them, we provided two practice signals with examples of "bad" captions that should receive low ratings. For instance, one practice trial presented a very weak, continuous vibration alongside captions that were clearly incorrect: sensory–*"It is very strong."*, emotional–*"It makes me feel excited."*, and associative–*"It reminds me of a heartbeat."* They received $15 USD in cash for their time, which exceeds the local minimum wage. For HapticCap, vibration sources and data collection procedure are detailed in (Hu et al., 2025).