

HapticCap: A Multimodal Dataset and Task for Understanding User Experience of Vibration Haptic Signals

Guimin Hu¹, Daniel Hershcovich¹, Hasti Seifi²

¹University of Copenhagen

²Arizona State University

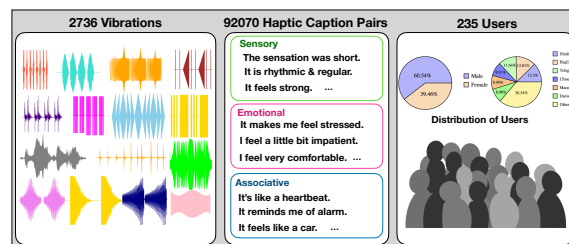
rice.hu.x@gmail.com, dh@di.ku.dk, hasti.seifi@asu.edu

Abstract

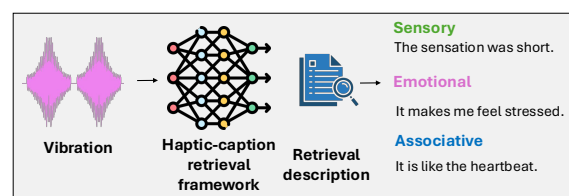
Haptic signals, from smartphone vibrations to virtual reality touch feedback, can effectively convey information and enhance realism, but designing signals that resonate meaningfully with users is challenging. To facilitate this, we introduce a multimodal dataset and task, of matching user descriptions to vibration haptic signals, and highlight two primary challenges: (1) lack of large haptic vibration datasets annotated with textual descriptions as collecting haptic descriptions is time-consuming, and (2) limited capability of existing tasks and models to describe vibration signals in text. To advance this area, we create *HapticCap*, the first fully human-annotated haptic-captioned dataset, containing 92,070 haptic-text pairs for user descriptions of sensory, emotional, and associative attributes of vibrations. Based on HapticCap, we propose the haptic-caption retrieval task and present the results of this task from a supervised contrastive learning framework that brings together text representations within specific categories and vibrations. Overall, the combination of language model T5 and audio model AST yields the best performance in the haptic-caption retrieval task, especially when separately trained for each description category. The dataset is available at <https://huggingface.co/datasets/GuiminHu/HapticCap>.

1 Introduction

Haptic signals can convey information and emotions through programmable feedback experienced by users via the sense of touch. From touchscreen and VR interactions (Choi et al., 2017, 2021) to gaming (Yun et al., 2023a), rehabilitation (Seim et al., 2022) and healthcare (Kuchenbecker et al., 2017; McMahan et al., 2011), haptic feedback has gained significant attention, with the market expected to exceed \$28 billion USD by 2026. However, little is known about how users perceive hap-



(a) Dataset with diverse vibrations, captions, and users.



(b) Retrieval task, addressing three description categories.

Figure 1: HapticCap dataset and haptic-caption retrieval task.

tic signals or how these signals relate to textual descriptions.

Existing multimodal datasets focus on modalities such as text, images, and audio, while overlooking vibration haptic signals, despite their significant potential for user applications (van der Linden et al., 2011; Seifi et al., 2020; Salvato et al., 2022). Language-based descriptions can express sensory (e.g., the intensity of tapping on a surface), emotional (e.g., the mood of a scene), and associative qualities (e.g., familiar feelings resembling real-world phenomena) of haptic signals (Seifi et al., 2015). These descriptions, which we refer to as *haptic captions*, are essential for designing vibrations with relevant sensory, emotional, and associative characteristics for users. In this regard, we first introduce the term haptic captions, along with the corresponding dataset and task (Figure 1) to refer to user descriptions of haptic signals.

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 made significant progress in recent years. In contrast to previous captioning research, haptic captioning is a largely unexplored area for two key reasons: (1) lack of large text-annotated vibration datasets covering sensory, emotional, and metaphorical associations; and (2) no available task and model to link vibration haptic signals with textual descriptions. Notably, the data collection process for haptic captioning is time-consuming and labor-intensive, further hindering progress in this domain.

In this paper, we introduce the first and largest fully human-annotated multimodal haptic captioning dataset for vibrations, HapticCap, which contains 92,070 vibration-description pairs with 2,736 unique vibration signals and 235 users and is collected over approximately 11 months¹. HapticCap captures diverse and well-differentiated haptic signals, along with rich textual descriptions that cover sensory, emotional, and associative perspectives. Building on this foundation, we propose a novel haptic-caption retrieval task that aims to retrieve textual descriptions from the perspectives of sensation, emotion, and metaphoric association for a given haptic signal. Inspired by captioning tasks in other modalities, such as image and audio (Mokady et al., 2021; Zhang et al., 2022), this task aims to bridge user language and vibration haptic signals. Given the scarcity of haptic-language data needed to train generative models, haptic caption generation is currently impractical. Instead, haptic caption retrieval task offers an interpretable framework to study semantic alignment between haptic signals and language: Given a vibration and a set of captions, can a model reliably identify the correct one?

We employ a supervised contrastive learning framework to project text and haptic representations into a shared space and align them with each other. We assess this framework’s performance by comparing models trained on individual description categories vs. all categories across various large models, showing the effectiveness of the proposed framework. The proposed task and framework for haptic caption retrieval provides several practical benefits: (a) it accelerates dataset expansion for generative models via human-in-the-loop annotation, where users quickly refine retrieved captions for new vibrations; (b) this task leverages vetted, human-curated captions to yield interpretable re-

sults, enabling applications such as accessibility, eyes-free notifications, and human-robot interaction in noisy environments; and (c) it objectively measures the alignment between haptic signals and captions, offering an evaluation framework for newly generated signals analogous to CLIP (Radford et al., 2021) in vision-language domains. As a pioneering effort in the haptic-text domain, our benchmark dataset and retrieval task establish a foundation for future progress in haptic language understanding. Our contributions include:

1. We introduce HapticCap, a novel human-annotated multimodal dataset with 92,070 vibration-description pairs. Each signal is annotated from three distinct perspectives: sensory, emotional, and associative, providing rich multidimensional descriptions. The dataset is available at <https://huggingface.co/datasets/GuiminHu/HapticCap>.
2. We propose a haptic-caption retrieval task to understand user experiences with vibration haptic signals and employ a contrastive learning framework to align and integrate haptic signals with textual descriptions.
3. We evaluate haptic-caption retrieval by comparing various models trained on combined and individual categories as baselines, paving the way for the development of sensory language models in haptics.

2 Related Work

2.1 Haptic Modality and User Experience

Vibrotactile technology is the most accessible and versatile form of haptics (Ege et al., 2011; Jung et al., 2024; García-Valle et al., 2017; Kaul and Rohs, 2017). Vibration signals can vary in amplitude and frequency over time, creating a large signal space to convey sensations and meanings (Seifi and MacLean, 2017). For instance, the VibViz library includes 120 vibrations with diverse signal and affective tags (Seifi et al., 2015). User language can enable designing meaningful haptic signals but previous work used preset lists of tags which are limited to the tags selected (Park et al., 2011; Israr et al., 2014a). Others focused on free-form natural language, but due to the time-consuming nature of data collection, they included small sets of 2-10 signals and reported qualitative trends in user descriptions (Obrist et al., 2013; Knibbe et al., 2018; Dalsgaard et al., 2022), preventing computational models for haptic descriptions.

¹<https://github.com/LeMei/HapticCap>.

2.2 Touch Datasets

Touch is one of the five basic senses, enabling humans and robots to perceive physical sensations. Recent touch datasets (Balasubramanian et al., 2024; Yang et al., 2022; Yuan et al., 2017) such as TVL (Fu et al., 2024), TLV (Cheng et al., 2024b), and Touch100k (Cheng et al., 2024a) consist of tactile images recorded as RGB data from a deformable sensor, resembling image or vision-based data. The tactile images in these datasets capture object shape, size, and texture (e.g., the tangible surface of a table) and are accompanied by text annotations generated partially or entirely by GPT-4V. These touch datasets enable training robots to perceive physical objects through tactile images. Unlike prior touch datasets, HapticCap focuses on vibration haptic signals—tactile feedback perceived by humans through handheld devices (e.g., mobile phones, VR controllers). These signals vary in amplitude and frequency over time and are inherently different from tactile images in prior datasets. HapticCap includes vibration signals with sensory, emotional, and associative descriptions which are fully annotated by humans. The dataset enables providing vibration feedback for users (rather than robots) interacting with digital or virtual content.

2.3 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. For haptics, Hu et al. (2024) developed a computational pipeline that extracts sensory and emotional tags from textual descriptions to link these keywords to haptic signal features, but this work focused on a small dataset consisting of 32 signals $\times$ 12 descriptions. Recently, Sung et al. (2025) developed a model that generates vibrations from user prompts; however, their work did not focus on captioning tasks and did not capture the sensory, emotional, and associative aspects of vibrations. In contrast, we propose a dataset and task to retrieve user captions for a haptic signal based on three descriptive categories.

3 HapticCap Dataset

We present our process for creating HapticCap over 11 months, which involves compiling diverse signals, collecting user descriptions, and performing manual checks and data validation to ensure the

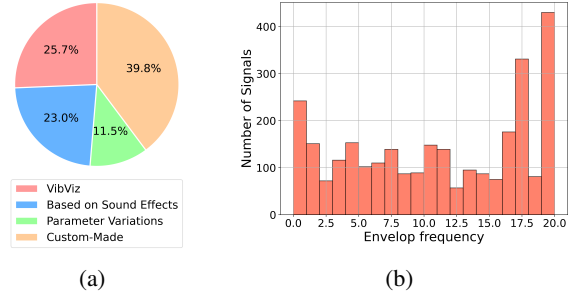


Figure 2: Illustration of haptic signal statistics: (a) distribution of haptic source, (b) signal count on envelope frequency feature.

dataset’s diversity and quality.

3.1 Source of Haptic Signals

We create a large, diverse set of 2,736 vibration signals by collecting an initial set of 304 diverse signals and further increasing signal variance by generating new signals. Figure 2 illustrates the sources of haptic signals and their distribution based on envelope frequency features, highlighting the diversity of both signal origins and characteristics.

The initial set of vibrations are from multiple sources to ensure variety: (a) we select 78 signals from VibViz (Seifi et al., 2015), focusing on signals that have distinct features and can render well on VR controllers, (b) we create 70 new vibrations based on sound effect libraries by either mimicking timing, or directly applying low-pass filtering to them (Ternes and MacLean, 2008; Degraen et al., 2021; Yun et al., 2023b), (c) we make 35 new vibrations by varying vibration parameters of amplitude, envelop frequency, carrier frequency, and rhythm (Yoo et al., 2015), (d) finally, we design 121 custom-made vibrations by applying various transformations to existing vibration signals such as time reversal (e.g., ramp up to ramp down), repeating part of the signal, and mixing subsets of vibrations to create new vibrations (Schneider and MacLean, 2016; MacLean et al., 2017). We normalize all vibration signals to 10 seconds by repeating the shorter signals. These methods are common practices for haptic designers to create new vibrations (Seifi et al., 2015; Ternes and MacLean, 2008; Yoo et al., 2015; Yun et al., 2023b).

Next, we follow procedures described by Lim et al. (2025) to perform stretching, amplifying, noise injection, and their combined operations on the haptic signals and generate 8 new haptic signals for each initial haptic signal. Lim et al. (2025)

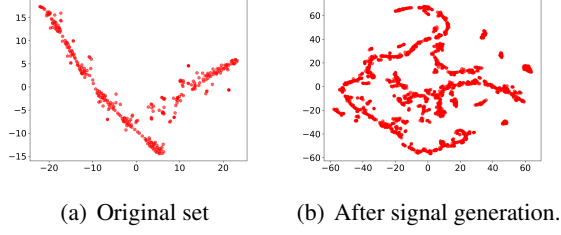


Figure 3: t-SNE visualization of haptic signal features (e.g., max, min).

shows that after these operations, vibrations will evoke similar haptic experience to the original signals. Also, these operations have been used in prior haptic literature to create new signals from existing ones (Israr et al., 2014b; Yoo et al., 2015; Israr et al., 2014b; Ternes and MacLean, 2008). This step increases the dataset size and diversity and improves the generalization of the haptic caption retrieval model. This generation resulted in a total of 2,736 unique vibration signals in our dataset. We ensure that each vibration haptic signal can be played on VR controllers (Meta Quest 3 and Pro) and have diverse waveform features (e.g., energy, frequency, change rate). We visualize haptic signals in feature space including the initial set and newly generated signals in Figure 3, further demonstrating our haptic signals exhibits both diversity and extensive spatial representation in feature space.

3.2 Collecting User Descriptions

We collect a haptic-caption dataset with 92,070 tuples of [vibration, description] after removing empty or NA response. Each vibration is described in English by at least 10 users according to sensory, emotional, or associative perspectives. Sensory refers to the participants’ senses (e.g., roughness), emotion involves positive or negative feelings (e.g., pleasant, agitating), and association reflects the real-world metaphors evoked by the signal (e.g., alarm clock).

We run a user study with 235 diverse users to collect descriptions of haptic signals. Collecting user descriptions for haptics is time-consuming as users need to feel the vibrations under controlled conditions in the lab. Participants wear headphones and earplugs to cancel any audio noise from VR controllers, following the established practice in haptic studies (Seifi et al., 2015; Yoo et al., 2015). They hold the VR controller in their non-dominant hand, feel each signal in a random order, and type three

descriptions about the sensory, emotional, and associative feel of the signal on a user interface. When they cannot think of a description for a signal, they enter Not Applicable (NA) in the response form. Each participant describes 16 haptic signals in one hour according to the three sensory, emotional, and associative aspects. Appendix C shows user backgrounds, example descriptions, and data collection platform.

3.3 Data Validation and Diversity Analysis

We perform three types of data analysis and validation, focusing on generated signals, agreement in captions, and diversity of signal-caption pairs.

First, we randomly sample 5% of the original haptic signals along with their corresponding generated versions to conduct human validation, assessing whether the haptic experience remains consistent—i.e., similar or different—between the original and generated signals. Each pair of signals is independently evaluated by three researchers outside the author team, using the following instruction: “If the two signals evoke similar sensory, emotional, and associative experiences, mark Y; otherwise, mark N”. As a result, 98.78% of the signals are verified to elicit a similar haptic experience to the original signals.

Second, human-labeled descriptions may contain errors, as annotators are inherently prone to mistakes such as misinterpreting instructions or accidental mislabeling (Weber-Genzel et al., 2024). This leads to high variance in human labeling across sensory, emotional, and associative dimensions. To address this, we consider that labels deviating significantly from the majority can be noisy or outliers. Thus, we calculate inter-annotator agreement scores based on sensory, emotional and associative descriptions and divide the dataset into low-agreement and medium/high-agreement subsets (Table 1). In the medium/high-agreement subset, we filter out HapticCap descriptions that have a low agreement with most other participants’ descriptions on a specific category. We first encode the description into a representation vector by T5 (Raffel et al., 2020) and then calculate cosine similarity for every pair of descriptions from different participants on the same category and the same haptic signal. Finally, we filter out descriptions with an average similarity score of less than 0.5 to other participants. Through this step, we primarily separate (1) the signals that may be less perceptible and their corresponding descriptions

	Full Dataset (Removing NA)	Medium/High Agreements
Vibration signals	2,736	2,709
Signal-sensory pairs	32,202	28,134
Signal-emotion pairs	30,762	25,092
Signal-association pairs	29,106	15,295
Total	92,070	68,521

Table 1: Statistics of HapticCap’s full dataset and medium/high-agreement subset.

and (2) descriptions that deviate significantly from the majority perception. We report the results of the haptic-caption retrieval task on both the full dataset and after removing the low-agreement data.

Third, we analyze the diversity of signal-caption pairs in the dataset (Figure 4). Specifically, we examine the distribution of haptic signals associated with emotion-related descriptions, as shown in Figure 4(a). Emotion categories (e.g., calm) are selected based on a word cloud generated from emotional aspects (Figure 7 in Appendix), and a signal is counted if it is reported to evoke the corresponding emotional experience. The distribution indicates the diversity of user experiences in the dataset. Similar analysis was performed for the sensory and associative categories in Figure 8 in Appendix. We further visualize the embedding spaces of captions in (Figure 4(b)) to show their diverse but also distinct characteristics spanning sensory, emotional, and associative dimensions. We provide further dataset analysis in Appendix B.

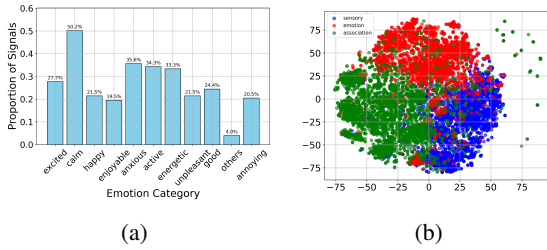


Figure 4: Diversity of signal-caption pairs in HapticCap: (a) emotion distribution of haptic signals, (b) t-SNE visualization of description embedding encoded by T5 (Raffel et al., 2020) on three aspects shown by three colors.

4 Haptic-Caption Retrieval Task

4.1 Task Formalization

Given a set of haptic signals $\mathcal{C} = \{h_1, \dots, h_n\}$, each signal is described by multiple participants across three dimensions: sensory, emotional, and associative qualities. Formally, a haptic signal h and a description $d^c \in \sum^*$ compose haptic-text

pair (h, d^c) , where $c \in \{s, e, a\}$ denotes sensory, emotional, or associative categories, respectively. The haptic-caption retrieval task’s objective is to retrieve the textual descriptions of three categories that correspond to a given haptic signal, using the haptic signal as the query and the descriptions as the target documents, as shown in Figure 5.

4.2 Encoders for Text and Haptics

Text representation: We evaluate text representations derived from three distinct language model architectures: encoder, encoder-decoder, and decoder. For this purpose, we test models representing each architecture type, including BERT-base (Devlin et al., 2019) (encoder), T5-base (Raffel et al., 2020) (encoder-decoder), Llama-3.2² (Touvron et al., 2023), and Mistral-7B-Instruct-v0.2 (Jiang et al., 2023) (both decoder). Specifically, we compute text representations by averaging the hidden states from the top layer of each model. BERT uses bidirectional context for tasks like question answering and classification, trained with Masked Language Modeling and Next Sentence Prediction. T5, an encoder-decoder transformer, frames tasks as sequence-to-sequence text generation. Llama is a decoder-only model focused on general-purpose language understanding, while Mistral uses a decoder-only architecture optimized for efficient natural language generation.

Haptic representation We test haptic representation by fine-tuning three audio models: AST (Gong et al., 2021), Wav2vec (Baevski et al., 2020), and EnCodec (Défossez et al., 2023). We adopted pre-trained audio models as haptic encoders for three main reasons. First, training a vibration model from scratch on HapticCap performs worse than fine-tuning audio models due to the size of the dataset (See Appendix A.1 for result). Second, vibration and audio signals are both one-dimensional signals and share key temporal and frequency-based characteristics, enabling the use of similar time-series analysis techniques. Third, prior work has shown perceptual similarities between the two modalities, with audio often used as a proxy in haptics research (Bernard et al., 2022; Pätzold et al., 2023; Yun et al., 2023b; Zhao et al., 2016; Degraen et al., 2021). Thus, we employ pre-trained audio models to encode haptic signals into representation. Note that we fine-tune the parameters of the pre-trained audio model to adapt it for haptic data encoding.

²meta-llama/Llama-3.2-3B-Instruct

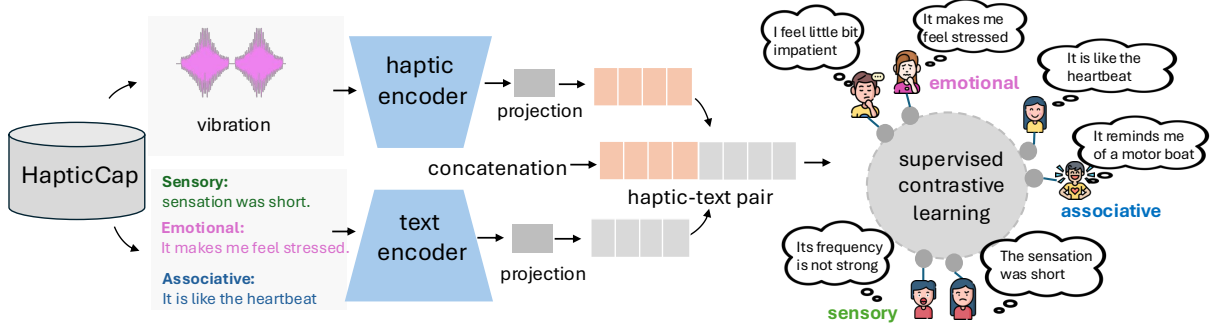


Figure 5: Overview of supervised contrastive learning for haptic signals and user descriptions. The proposed framework consists of pairing signals with their descriptions for each category, aiming to pull pairs with the same label closer and pull those with different labels away in the embedding space. The figure shows the architecture of the model trained on the sensory, emotional, and associative categories.

AST converts audio into 128-dimensional log Mel filterbank features and uses the spectrogram as input. Wav2Vec learns representations from unlabeled audio by predicting the next frame, using contrastive learning to capture speech patterns. EnCodec compresses raw audio into a compact latent representation, which is then stored, transmitted, and reconstructed.

4.3 Supervised Contrastive Learning Framework

Let $f_H : \mathcal{R}^{T \times m} \rightarrow \mathcal{R}^{d_1}$ be a haptic signal encoding function (e.g., a pre-trained audio model), which takes an m -dimensional haptic signal of length up to T and outputs a d_1 -dimensional vector as the representation of haptic signal. Let $f_D : \Sigma^* \rightarrow \mathcal{R}^{d_2}$ be a text encoding function (e.g., from a pretrained language model) that produces a d_2 -dimensional vector representation of a sentence. Note that we fine-tune the parameters of the pre-trained audio model to better adapt it for haptic data encoding.

Supervised contrastive learning aims to pull the clusters of points belonging to the same class together in an embedding space and simultaneously pushes apart clusters of samples from different classes (Khosla et al., 2020). We initialize the parameters of the haptic encoder and fine-tune the last n layers of the text encoder. After the haptic and text encoders, we set up two linear projection layers, one for each representation, aiming to map both representations into a common feature space with a dimension of d . First, we pair each haptic representation with its corresponding description to create a haptic-text pair. Next, we concatenate the haptic ID (e.g., F1) and category (e.g., sensory) as the label of haptic-text pair (e.g., F1_sensory). In this setting, the haptic-text pairs from multiple

participants but for the same haptic signal and same category can be viewed as the same class. Specifically, each haptic signal has multiple labels from different users, and the descriptions are categorized into three groups: sensory, emotional, and associative. To address this, we combine the category and haptic ID as the label for supervised contrastive learning, with the goal of linking the haptic signal to its descriptions within a specific category (Figure 5). We define the contrastive loss as:

$$\mathcal{L}^{\text{sup}} = \sum_{i \in \mathcal{I}} \frac{-1}{|P(i)|} \sum_{j \in P(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_j / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)} \quad (1)$$

Here, $P(i)$ is the set of indices of all positives in the batch distinct from anchor haptic-text pair i , and $|P(i)|$ is its cardinality. $A(i) = \{j \in P(i) : y_i = y_j\}$, where y_i and y_j represent the labels of haptic-text pairs i and j , respectively. The condition $y_i = y_j$ indicates that the two pairs share the same haptic signal and their descriptions belong to the same category. $\mathbf{z}_*$ represents the haptic-text pair representation obtained by concatenating the haptic and text representations. τ represents the temperature coefficient.

Suppose we partition haptic-text pair set $\mathcal{C}$ into mutually exclusive train, validation, and test sets, $\mathcal{C} = \mathcal{C}_{\text{train}} \cup \mathcal{C}_{\text{valid}} \cup \mathcal{C}_{\text{test}}$. In practice, we first divide all haptic signals into mutually exclusive training, validation, and test sets. Captions are then associated with these sets based on the underlying haptic signals, ensuring that there is no overlap of signals between the splits. We train contrastive learning framework via minimizing training loss, leverage the validation set to optimize the selection of hyperparameters within the framework, and evaluate model performance on $\mathcal{C}_{\text{test}}$ for the haptic-caption retrieval task. For $h \in \mathcal{C}_{\text{test}}$:

$$d = \underset{d \in C_{test}}{\text{top-K}} (\text{sim}(\text{text}(d), \text{haptic}(h))) \quad (2)$$

where $d = \{d^s, d^e, d^a\}$ is the retrieved haptic descriptions of all three categories corresponding to a given haptic signal, and all three categories are candidates for retrieval. Prediction is counted as correct if the retrieved description is in the set of the top K most similar objects to a given haptic signal in semantics. We use the following similarity function between the textual description and haptic representations based on cosine similarity,

$$\text{sim}(\text{text}(d), \text{haptic}(h)) = \text{softmax}(\kappa \cdot \langle f_D(d), f_H(h) \rangle) \quad (3)$$

where $\langle, \rangle$ denotes the dot product, softmax is softmax function, κ is scaling factor to adjust the range of similarity scores.

5 Experiments

5.1 Setup

We evaluate the proposed contrastive learning framework from two perspectives: (1) performance on combined categories i.e., the framework is trained using all sensory, emotional, and associative categories, and evaluated on both combined and individual categories, as shown in Section 6.1, and (2) performance on individual categories i.e., the framework is independently trained and evaluated for each category, as discussed in Section 6.2. For both perspectives, we report results on the full dataset and after removing the low-agreement subset of the data. Additionally, we report the results of a zero-shot generalization test on our proposed framework in Table 5 in the Appendix.

5.2 Evaluation Metrics

We use common metrics in Information Retrieval (IR), including Precision@10 (P@10), Recall@10 (R@10), mean Average Precision at 10 (mAP@10), and normalized Discounted Cumulative Gain at 10 (nDCG@10). Precision ensures that the most relevant descriptions appear at the top, while recall is essential for capturing all relevant options. mAP@10 provides a single score that takes both the precision and rank of the retrieved descriptions into account. nDCG@10 gives higher weight to relevant items that appear earlier in the top 10 results and normalizes the weighted score to be between 0 and 1. Together, these comprehensive metrics provide valuable benchmarks for various downstream applications in haptic caption retrieval.

5.3 Implementation Details

All experiments run with one NVIDIA RTX A100 GPU. We perform a light grid search over the hyperparameters, learning rates $\alpha \in \{10^{-3}, 10^{-4}, 10^{-5}\}$, temperature coefficients $\tau \in \{0.07, 0.1\}$, $n = \{1, 2, 3, 4, 5\}$, and $m = \{1, 2, 3, 4, 5\}$. The model achieves optimal performance when $\alpha = 10^{-3}$, $\tau = 0.1$, $n = 3$ and $m = 2$ for best combination of T5 and AST. We set $K = 10$, $\kappa = 100$, $d_1 = 768$, $d_2 = 768$, and common dimension $d = 768$. The batch size is set to 128, and training is conducted for 15 epochs, taking approximately 4 to 12 hours. The optimal hyperparameters with the best epoch for early stopping are determined using a validation set.

We initialize the parameters of haptic encoder (audio model) and fine-tune last n and m layers of text encoder (LLM) and haptic encoder, respectively (See hyperparameter analysis in Appendix A.4). In our proposed model, the trainable parameters consist of the last n layers of the text encoder and last m layers of haptic encoder, and two linear projection layers. The data is divided into 70% training, 10% validation, and 20% test subsets for the full dataset and the medium/high agreement subset (See Appendix A.2 for more implementation details).

6 Results

6.1 Performance of Model Trained on Combined Categories

Table 2 presents the performance of the proposed framework trained in all categories combined, utilizing separately four text model architectures with three fine-tuned audio models. Also, we report the performance of the model trained on all categories when evaluated separately for the sensory, emotion, and association categories. The chance-level performance for the three metrics (P@10, R@10, MAP@10) is 7.81%, estimated based on randomly selecting 10 captions from a pool of 128, with 10 ground-truth captions.

First, the model after filtering out low-agreement data performs better, improved by average around 20% on all metrics, indicating filtering is effective and helpful to performance (See Appendix A.4 for results obtained with alternative filtering thresholds). Second, our model performs best on the emotional category, followed by performance in sensory category, with the poorest performance in associative category across combinations of language

		P@10	R@10	mAP@10	nDCG@10	Sensory	Emotional	Associative
BERT	AST	13.68/12.01	16.09/13.31	27.30/23.89	0.4952/0.4514	13.26/11.42	14.21/12.96	12.69/11.04
	WavVec	11.75/10.96	15.78/13.16	26.58/23.16	0.4916/0.4487	11.41/10.59	13.69/12.01	11.46/10.25
	EnCodec	12.64/10.68	16.10/13.30	27.42/23.75	0.4956/0.4574	11.48/10.08	12.89/11.76	11.36/10.07
T5	AST	16.66*/12.25*	19.14*/13.19*	30.62*/24.03*	0.5579*/0.4610*	16.85*/11.51*	17.30*/12.98*	15.16*/11.20*
	WavVec	16.01/10.90	18.86/13.27	29.48/23.97	0.5368/0.4563	15.95/11.08	16.89/12.67	12.09/10.21
	EnCodec	15.96/10.98	18.87/13.10	29.32/23.85	0.5316/0.4563	15.03/11.19	16.18/12.46	12.48/10.32
Mistral	AST	16.03/12.36	18.58/13.60	29.92/23.91	0.5527/0.46002	16.29/11.53	16.89/12.87	15.09/11.09
	WavVec	15.89/10.68	18.42/13.28	29.27/23.44	0.5484/0.4474	15.31/10.47	15.93/12.66	12.12/10.48
	EnCodec	15.28/10.49	17.51/13.11	28.68/23.26	0.5334/0.4460	15.11/10.39	15.63/12.62	12.36/9.91
Llama	AST	16.54*/12.37*	19.03*/13.67*	30.47*/24.32*	0.5536*/0.4731*	16.59*/11.60*	17.28*/13.04*	15.04*/11.28*
	WavVec	16.14/11.11	18.36/13.59	29.26/24.30	0.5488/0.4689	16.08/11.67	16.89/13.10	12.18/10.29
	EnCodec	15.38/10.72	15.62/13.60	25.93/24.51	0.4880/0.4690	16.33/10.85	16.40/12.62	11.66/10.29

Table 2: Performance evaluation of framework trained on all categories (sensory, emotional, and associative). The three right columns show P@10 of the framework trained on all three categories but tested on individual categories. For each metric, the numbers before and after “-/-” indicate the performance results obtained after filtering out low-agreement data and performance on the full dataset, respectively. The **best** and **second-best** results are highlighted with green and orange. The chance-level performance for P@10, R@10, and MAP@10 is 7.81%. Significant differences of $p < 0.05$ using t-test are indicated by *.

		Sensory				Emotional				Associative			
		P@10	R@10	mAP@10	nDCG@10	P@10	R@10	mAP@10	nDCG@10	P@10	R@10	mAP@10	nDCG@10
BERT	AST	15.17/12.65	15.86/14.18	27.62/24.30	0.5116/0.4709	15.62/12.92	17.80/14.22	28.13/25.66	0.5138/0.4736	14.26/11.89	15.20/14.10	27.32/24.01	0.5022/0.4682
	WavVec	14.39/11.43	15.10/13.89	26.84/23.75	0.5033/0.4601	14.49/11.16	15.89/	27.96/24.15	0.5018/0.4605	12.11/10.88	14.39/13.04	26.19/23.26	0.4928/0.4601
	EnCodec	14.63/11.37	15.25/13.61	27.10/23.12	0.5035/0.4536	15.48/11.47	17.88/13.48	27.16/24.37	0.5032/0.4625	13.08/10.22	14.57/	26.69/23.68	0.4933/0.4644
T5	AST	17.24*/13.48	17.65*/14.61	30.17*/25.58	0.5266*/0.4779	18.89*/13.74	20.79*/15.62	30.56*/27.10	0.5274*/0.4886	17.88*/12.54	20.21*/15.06	29.36*/24.57	0.5248*/0.4751
	WavVec	16.61/12.87	16.34/13.66	29.87/24.85	0.5078/0.4648	17.60/12.20	18.99/14.75	27.24/26.14	0.5016/0.4791	15.99/11.40	17.12/14.10	28.33/23.49	0.5087/0.4648
	EnCodec	16.47/12.57	18.48/13.43	29.24/24.48	0.5025/0.4611	16.28/11.99	17.45/14.38	26.84/26.04	0.4996/0.4725	15.58/11.11	17.01/13.74	27.80/23.20	0.5034/0.4613
Mistral	AST	16.62/12.87	17.64*/14.48	29.28/24.37	0.5137/0.4748	16.86/13.07	19.34/15.37	28.67/26.04	0.5108/0.4743	16.83/11.08	20.07*/14.09	28.55/23.20	0.5156/0.4610
	WavVec	16.38/12.36	17.16/14.11	28.52/24.05	0.5073/0.4714	16.04/12.84	16.29/15.03	27.01/25.68	0.4982/0.4710	15.61/10.46	18.14/13.70	27.79/23.04	0.4940/0.4557
	EnCodec	16.48/12.18	17.33/13.89	28.42/23.78	0.5106/0.4709	16.38/12.67	17.84/14.78	28.25/25.61	0.5015/0.4705	15.76/10.44	19.11/13.76	27.37/24.12	0.5114/0.4547
Llama	AST	17.18*/13.28	17.47/14.50	29.95*/25.21	0.5211*/0.4718	18.48*/13.48	20.38*/15.21	30.31*/26.85	0.5217*/0.4852	17.82*/12.15	20.05*/14.97	29.12*/24.18	0.5226*/0.4709
	WavVec	16.29/13.07	16.20/13.35	26.52/24.14	0.5099/0.4687	16.52/13.11	19.26/14.83	27.02/26.33	0.5101/0.4810	16.98/12.02	20.01/24.67	28.79/23.86	0.5118/0.4678
	EnCodec	16.67/13.14	16.33/13.38	29.21/24.17	0.5111/0.4654	16.69/13.13	19.31/14.67	29.27/26.21	0.5124/0.4812	17.28/12.04	20.83/20.56	30.86/24.08	0.5140/0.4675

Table 3: Performance evaluation of framework trained on individual category. The numbers before and after the “-/-” denote results on the medium/high agreement subset (i.e., after excluding low-agreement data) and the full dataset, respectively. The chance-level performance for P@10, R@10, and MAP@10 is 7.81%. Significant differences using t-tests are marked by *.

models (e.g., BERT, T5, Mistral, and LLaMA) and AST. We speculate that higher agreement enhances performance, as emotional descriptions generally yield higher agreement, while associative descriptions tend to yield lower agreement. Third, the combination of T5 and AST, as well as Llama and AST, shows superior overall performance across multiple metrics. Specifically, T5 with AST achieves the highest P@10 (16.66), R@10 (19.14), mAP@10 (30.62), and nDCG@10 (0.5579) across all three categories. Llama with AST closely follows in all categories and metrics. In summary, the proposed contrastive learning framework aligns haptic signals and their textual descriptions. For modal encoder, BERT lags behind T5 and Llama in the haptic caption retrieval task, showing certain limitations. T5 excels in natural language understanding tasks, while Llama 3.2 performs strongly in natural language generation (Fu et al., 2023), leading to T5 slightly outperforming Llama. Regarding haptic en-

coder, AST is an effective haptic feature extractor and can be paired with LLMs (e.g., T5 and Llama) for the haptic caption task. WavVec is trained on speech corpora, while EnCodec is designed to restore the original signal after compression. See the lower performance of vibration-specific encoders trained on HapticCap (Appendix A.1).

6.2 Performance of Model Trained on Individual Category

We also evaluate the performance of the proposed framework trained in a single category (i.e., sensory, emotional, and associative) across different combinations of pretrained language models and pre-trained audio models. The results are presented in Table 3. Among the models, T5 with AST consistently outperforms others across multiple metrics, particularly excelling in all three categories with the highest scores across the board. In comparison, Llama with AST shows the second-best per-

formance on most metrics. As exceptions, Mistral with AST outperforms Llama with AST in R@10 for the sensory and associative categories. The performance of T5+AST and Llama+AST aligns with the results obtained from training on all categories. To provide further insights, we include concrete examples of successful and failed retrievals for T5+AST in Appendix B.2. Furthermore, we conduct zero-shot generalization experiments and report the results in Appendix A.3.

We attribute AST’s strong performance in the combined and category-specific results to two factors: (1) Model architecture: AST treats spectrogram patches as tokens and models their relationships with a Transformer, capturing both global patterns (e.g., rhythm, frequency shifts) and local textures important for haptic perception (Abou Chahine et al., 2022; Ternes and MacLean, 2008). In contrast, EnCodec and Wave2Vec operate directly on raw signals: EnCodec emphasizes local patterns for compression (Défossez et al., 2023), while Wave2Vec learns frame-level speech dependencies, making both less suited to vibration dynamics. (2) Training data: AST is pretrained on AudioSet (Gemmeke et al., 2017), which covers diverse everyday sounds with temporal and frequency structures relevant to haptics (e.g., object collisions, surface interactions) (Lee and Choi, 2013; Kim et al., 2023). In contrast, Wave2Vec is trained mainly on speech, and EnCodec is optimized for speech and music, limiting their generalization to vibration.

Conclusion

In this paper, we introduce the haptic-caption retrieval task to link user descriptions to vibration haptic signals. To support this task, we present a large haptic caption dataset called HapticCap, carefully designed and validated to ensure diversity and quality. Our primary contribution lies in introducing a novel dataset and task for the community, and demonstrating its feasibility for haptic-language understanding. We also report results from a supervised contrastive learning framework that incorporates various pre-trained models for text and haptic, aligning user descriptions with vibration signals. Since haptic-text alignment is a nascent area with limited prior work, we intentionally adopt contrastive learning as a proven method to provide a strong and interpretable baseline for evaluating the HapticCap dataset. Our results suggest that

haptic-caption retrieval is a challenging task for existing pre-trained models. Our experimental results and dataset provide a baseline for future work in this area and pave the way for the development of sensory language models for haptics. We are also working on haptic captions with an end-to-end generative framework that takes vibration signal as input and output corresponding captions, aiming to further contribute to the NLP and Haptics communities in our future work.

Limitations

First, HapticCap is fully manually annotated through a time-intensive process. While it represents the largest and most diverse collection of vibration signal-caption pairs to date, it may still not encompass the full range of possible signals and user experiences. Additionally, user descriptions are only available in English, and data collection is restricted to Denmark and the United States, potentially introducing bias due to geographical limitations. Second, we acknowledge that there is some semantic overlap across categories. For example, a sensory description “*It’s a funny buzz, like a childish tingle*” may also convey emotional and associative meanings. Such overlaps may explain why training a combined model to push away categories sometimes yields worse performance compared to treating categories independently. Future work can model the interdependencies among description categories to enhance haptic caption retrieval performance. Third, the contrastive learning framework we adopt is a well-established method for aligning multiple modalities and has shown strong performance on various multimodal tasks. As the first work to explore the integration of language and vibration, leveraging this framework offers a solid foundation for cross-modal alignment. Yet, its maturity limits the novelty of our approach.

Ethical Considerations

The collection and management of personal data in our study were approved by the university’s Institutional Review Board (IRB). Additionally, the collection of user information and haptic experience data was conducted with the consent of participants. Participants used pseudonyms, and no identifying information was included in the dataset. We reviewed the haptic descriptions to ensure they do not contain any offensive content.

Acknowledgement

We sincerely thank Pranav Iyer for his invaluable efforts in vibration signal design and data collection. We also thank the study participants for providing high-quality data annotations. This work was supported by research grants from VILLUM FONDEN (VIL50296) and the National Science Foundation (#2339707).

References

- Ramzi Abou Chahine, Dongjae Kwon, Chungman Lim, Gunhyuk Park, and Hasti Seifi. 2022. Vibrotactile similarity perception in crowdsourced and lab studies. In *International Conference on Human Haptic Sensing and Touch Enabled Computer Applications*, pages 255–263. Springer.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. [wav2vec 2.0: A framework for self-supervised learning of speech representations](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Jagan K Balasubramanian, Bence L Kodak, and Yasemin Vardar. 2024. Sens3: Multisensory database of finger-surface interactions and corresponding sensations. In *International Conference on Human Haptic Sensing and Touch Enabled Computer Applications*, pages 262–277. Springer.
- Corentin Bernard, Jocelyn Monnoyer, Michaël Wiertelowski, 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, Heather Culbertson, Mark R. Miller, Alex Olwal, and Sean Follmer. 2017. [Grability: A Wearable Haptic Interface for Simulating Weight and Grasping in Virtual Reality](#). In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*, UIST ’17, pages 119–130, New York, NY, USA. Association for Computing Machinery.
- 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.
- 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.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics.
- Enes Selman Ege, Furkan Cetin, and Cagatay Basdogan. 2011. Vibrotactile feedback in steering wheel reduces navigation errors during gps-guided car driving. In *IEEE World Haptics Conference (WHC)*, pages 345–348. IEEE.
- 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*.
- Zihao Fu, Wai Lam, Qian Yu, Anthony Man-Cho So, Shengding Hu, Zhiyuan Liu, and Nigel Collier. 2023. Decoder-only or encoder-decoder? interpreting language model as a regularized encoder-decoder. *arXiv preprint arXiv:2304.04052*.
- Gonzalo García-Valle, Manuel Ferre, Jose Breñosa, and David Vargas. 2017. Evaluation of presence in virtual environments: haptic vest and user’s haptic skills. *IEEE Access*, 6:7224–7233.
- Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.
- Yuan Gong, Yu-An Chung, and James R. Glass. 2021. [AST: audio spectrogram transformer](#). In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia*,

- August 30 - September 3, 2021, pages 571–575. ISCA.
- 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, Siyan Zhao, Kaitlyn Schwalje, Roberta Klatzky, and Jill Lehman. 2014a. Feel Effects: Enriching Storytelling With Haptic Feedback. *ACM Transactions on Applied Perception (TAP)*, 11(3):1–17. Publisher: ACM New York, NY, USA.
- Ali Israr, Siyan Zhao, Kaitlyn Schwalje, Roberta Klatzky, and Jill Lehman. 2014b. Feel effects: enriching storytelling with haptic feedback. *ACM Transactions on Applied Perception (TAP)*, 11(3):1–17.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *CoRR*, abs/2310.06825.
- Kyungeun Jung, Sangpil Kim, Seungjae Oh, and Sang Ho Yoon. 2024. Hapmotion: motion-to-tactile framework with wearable haptic devices for immersive vr performance experience. *Virtual Reality*, 28(1):13.
- Oliver Beren Kaul and Michael Rohs. 2017. Haptic-head: A spherical vibrotactile grid around the head for 3d guidance in virtual and augmented reality. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, pages 3729–3740.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.
- Dong-Geun Kim, Jungeun Lee, Gyeore Yun, Hong Z Tan, and Seungmoon Choi. 2023. Sound-to-touch crossmodal pitch matching for short sounds. *IEEE Transactions on Haptics*, 17(1):2–7.
- Jarrold Knibbe, Adrian Alsmith, and Kasper Hornb  k. 2018. [Experiencing Electrical Muscle Stimulation](#). *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–14. Publisher: ACM New York, NY, USA.
- Katherine J. Kuchenbecker, Robert C. Parajon, and Margrit P. Maggio. 2017. [Evaluation of a Vibrotactile Simulator for Dental Caries Detection](#). *Simulation in Healthcare*, 12(3):148.
- Jaebong Lee and Seungmoon Choi. 2013. Real-time perception-level translation from audio signals to vibrotactile effects. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2567–2576.
- Chungman Lim, Gyeongdeok Kim, Su-Yeon Kang, Hasti Seifi, and Gunhyuk Park. 2025. Can a machine feel vibrations?: Predicting roughness and emotional responses to vibration tactions via a neural network. *IEEE Transactions on Haptics*.
- 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*.
- 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.
- William McMahan, Jamie Gewirtz, Dorsey Standish, Paul Martin, Jacquelyn A. Kunkel, Magalie Lilavois, Alexei Wedmid, David I. Lee, and Katherine J. Kuchenbecker. 2011. [Tool Contact Acceleration Feedback for Telerobotic Surgery](#). *IEEE Transactions on Haptics*, 4(3):210–220. Conference Name: IEEE Transactions on Haptics.
- Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.
- 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.
- Gunhyuk Park, Seungmoon Choi, Kyunghun Hwang, Sunwook Kim, Jaecheon Sa, and Moonchae Joung. 2011. Tactile Effect Design and Evaluation for Virtual Buttons on a Mobile Device Touchscreen. In *Proceedings of the International Conference on Human Computer Interaction with Mobile Devices and Services*, pages 11–20.
- Bastian P  tzold, 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.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *J. Mach. Learn. Res.*, 21:140:1–140:67.
- Mike Salvato, Sophia R. Williams, Cara M. Nunez, Xin Zhu, Ali Israr, Frances Lau, Keith Klumb, Freddy Abnoui, Allison M. Okamura, and Heather Culbertson. 2022. [Data-Driven Sparse Skin Stimulation Can Convey Social Touch Information to Humans](#). *IEEE Transactions on Haptics*, 15(2):392–404.
- 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, Matthew Chun, Colin Gallacher, Oliver Schneider, and Karon E MacLean. 2020. How do novice hapticians design? a case study in creating haptic learning environments. *IEEE transactions on haptics*, 13(4):791–805.
- 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.
- 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.
- Janet van der Linden, Rose Johnson, Jon Bird, Yvonne Rogers, and Erwin Schoonderwaldt. 2011. [Buzzing to Play: Lessons Learned from an in The Wild Study of Real-Time Vibrotactile Feedback](#). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’11, pages 533–542, New York, NY, USA. Association for Computing Machinery.
- 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.
- Leon Weber-Genzel, Siyao Peng, Marie-Catherine de Marneffe, and Barbara Plank. 2024. Varierr nli: Separating annotation error from human label variation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2256–2269.
- 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.
- Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. 2022. Touch and go: Learning from human-collected vision and touch. *arXiv preprint arXiv:2211.12498*.
- 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.
- Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. 2017. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762.
- Gyeong 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.
- Gyeong 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.
- Yiming Zhang, Hong Yu, Ruoyi Du, Zhanyu Ma, and Yuan Dong. 2022. Caption feature space regularization for audio captioning. *arXiv preprint arXiv:2204.08409*.
- Siyan Zhao, Zachary Schwemler, Adam Fritz, and Ali Israr. 2016. Stereo haptics: Designing haptic interactions using audio tools. In *Proceedings of the TEI’16*:

Tenth International Conference on Tangible, Embedded, and Embodied Interaction, pages 778–781.

A Additional Experiments

A.1 Pretraining a Haptic Encoder

We also train a haptic-specific encoder from scratch using our vibration signals, with AST and Encodec architectures, respectively. Furthermore, we integrate this encoder with T5 and LLaMA in our contrastive learning framework and evaluate their performance on the haptic caption retrieval task, as shown in Table 4.

A.2 More Implementation Details

	All Categories	Sensory	Emotion	Association
Ours	16.66	16.85	17.30	15.16
T5+AST	14.19↓	14.68↓	15.12↓	11.53↓
Llama+AST	14.37↓	14.74↓	15.26↓	11.61↓
T5+Encodec	12.67↓	13.19↓	14.06↓	10.44↓
Llama+Encodec	13.25↓	13.43↓	13.71↓	10.26↓

Table 4: Results on pretraining audio models (e.g., AST) with haptic signals and integrating with large language models (e.g., T5, Llama) on P@10 metric.

We present the hyperparameter setting for the combinations of text encoder (e.g., LLMs T5 and Llama) and haptic encoder (e.g., AST, WavVec and Encodec) as Table 8. We fuse haptic and text representations using simple concatenation, a widely used and straightforward feature fusion approach. As the first work in this area, our primary goal is to evaluate the performance of recent LLMs and audio models on the proposed HapticCap dataset, thus we did not include results on various fusion strategies in this work.

Compared to fine-tuning pretrained audio models, the performance of the model with a haptic-specific encoder performance drops notably. This is perhaps because the fine-tuned audio models better learn temporal and rhythmic features, which are shared between vibrations and audio (Bernard et al., 2022), from much larger audio datasets than HapticCap, and the fine-tuning further adapts them to haptic signals.

A.3 Zero-Shot Generalization

A zero-shot generalization test assesses a model’s ability to handle tasks or data it has never encountered during training. In our work, we evaluate the performance of the proposed framework in zero-shot generalization. Specifically, we test the specific-category model’s performance on unseen categories. For instance, we train the framework with the sensory description and test it on the emo-

tional and associative categories. The results are presented in Table 5.

First, models perform best when tested on the same category they were trained on (diagonal values). When tested on other categories, performance drops, indicating a domain shift in description styles. For example, Sensory-trained models struggle with emotional and associative captions. Second, Sensory-trained models perform better across different categories compared to emotion-trained and associative-trained models. Sensory-related descriptions can overlap with both emotional (e.g., “intense” can convey strong emotions) and associative (e.g., “beat” or “pulse” are related to heartbeat). In contrast, emotional and associative categories are more domain-specific, making cross-category adaptation more challenging. These results suggest that feature alignment methods, such as adversarial domain adaptation, could help mitigate the gap between different description styles. In terms of generalization, sensory training provides better generalization than emotional or associative training.

A.4 Hyperparameter Analysis

We also conduct experiments with different values for the number of fine-tuning last layer n of text encoder T5. The model achieves the optimal result when $n=3$, both for combined and separate categories. The detailed P@10 results for $n \in \{1, 2, 3, 4, 5\}$ are reported in Table 7.

For the filtering threshold, the choice of 0.5 was based on empirical evaluation to balance data quality and quantity by excluding inconsistent descriptions while retaining sufficient training samples. We examined alternative thresholds of 0.4 and 0.6 in Table 6. A lower threshold (0.4) tends to retain more noisy captions, while a higher threshold (0.6) removes more data, limiting training examples.

B Dataset Analysis

B.1 Diversity Analysis

We perform a comprehensive diversity analysis of the proposed HapticCap dataset. Figure 6 presents the distribution of vibration haptic signal based on amplitude and zero crossing rate features. Figure 7 shows word-cloud visualizations corresponding to the sensory, emotional, and associative captions. Additionally, Figure 8 illustrates the distribution of haptic signals associated with sensory and associative words, which are extracted from their respective word clouds.

		Sensory				Emotional				Associative			
		P@10	R@10	mAP@10	nDCG@10	P@10	R@10	mAP@10	nDCG@10	P@10	R@10	mAP@10	nDCG@10
T5+AST	Sensory	18.16	20.23	31.32	0.5318	12.47	15.52	23.57	0.4607	11.71	13.22	20.65	0.4461
	Emotion	11.65	13.10	21.09	0.4467	17.45	20.38	29.56	0.5298	11.01	13.08	19.25	0.4381
	Association	10.10	11.89	20.56	0.4241	10.08	11.69	20.81	0.4309	17.01	20.36	29.49	0.5225
Llama+AST	Sensory	18.19	19.31	28.68	0.5208	12.19	15.49	23.24	0.4593	11.68	13.14	20.36	0.4427
	Emotion	11.24	13.05	20.88	0.4416	14.75	17.16	28.36	0.5219	11.05	13.41	18.61	0.4312
	Association	10.07	11.34	20.22	0.4420	10.10	10.86	20.08	0.4326	18.16	22.32	29.64	0.5109

Table 5: The performance of zero-shot generation. Row headers denote the combination of language and haptic encoders and the description category used for model training, and column headers denote the category used for testing the model without fine-tuning.

Threshold	P@10	R@10	mAP@10	nDCG@10	Sensory	Emotional	Associative
threshold=0.4	15.57	18.41	29.30	0.5343	15.91	16.92	14.10
threshold=0.5 (ours)	16.66	19.14	30.62	0.5579	16.85	17.30	15.16
threshold=0.6	15.61	18.22	29.16	0.5375	15.89	16.84	14.47

Table 6: Performance comparison across different thresholds for filtering out low-agreement data.

	All Categories	Sensory	Emotion	Association
n=1	11.47↓	10.78↓	11.91↓	10.36↓
n=2	15.61↓	15.36↓	16.82↓	15.19↓
n=3	16.74↓	16.59↓	17.28↓	15.44↓
n=4	14.94↓	14.51↓	15.64↓	14.25↓
n=5	14.39↓	14.16↓	14.86↓	14.11↓

Table 7: Hyperparameter analysis on fine-tuning layer on P@10 metric.

	n	m	α	τ
T5 + AST	3	2	10^{-3}	0.1
T5 + WavVec	3	3	10^{-3}	0.1
T5 + Encodec	3	3	10^{-3}	0.1
Llama + AST	2	2	10^{-3}	0.1
Llama + WavVec	2	3	10^{-3}	0.1
Llama + Encodec	2	3	10^{-3}	0.1

Table 8: The values of hyperparameters.

Regarding sensory aspect, words such as short, long, strong, beat, pulse, fast, and interval appear prominently, indicating that sensory descriptions focus on measurable and temporal characteristics of haptic signals. These terms suggest that users describe touch stimuli based on rhythm, intensity, and speed. For emotional category, the most dominant words include calm, energetic, anxious, uncomfortable, excited, good, and happy, demonstrating that haptic signals evoke clear emotional responses. The presence of both positive (e.g., calm, energetic) and negative (e.g., anxious, uncomfortable) emotions indicates a diverse range of affective reactions. In term of associative, words like alarm, car, phone, music, game, machine, and heartbeat dominate this

category, reflecting the tendency of users to associate haptic sensations with real-world experiences, objects, or events. These associations suggest that haptic signals can be linked to familiar contexts, making them more interpretable.

Overall, the word clouds highlight the diversity of user experiences and the distinct focus of each aspect: sensory descriptions emphasize physical properties, emotional descriptions capture feelings, and associative descriptions link haptic experiences to real-world references.

The distribution of signals over sensory and associative words in Figure 8 further suggest the vibration signals in HapticCap can create diverse experiences (see Figure 4a for the emotional category). Finally, we calculate distinct-n and n-gram of description to measure the diversity of textual description, shown in Table 9.

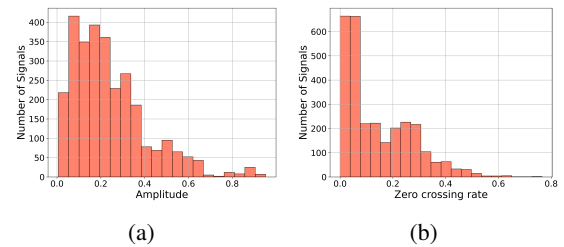
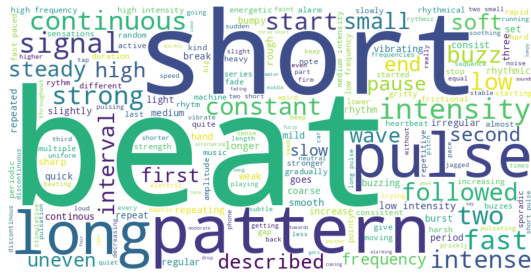


Figure 6: Haptic signal statistics : (a) signal count on amplitude feature, (b) signal count on zero crossing rate.

B.2 Case Analysis

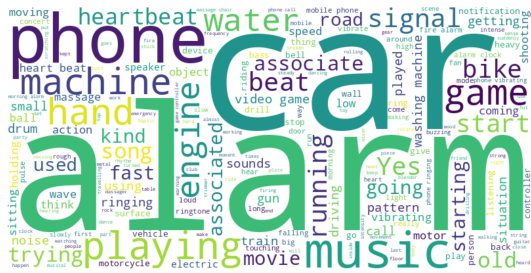
Table 10 shows the top-10 results for a selected haptic signal in the sensory category for T5 + AST, with correct retrievals in bold font.



(a) Sensory



(b) Emotional



(c) *Associative*

Figure 7: Visualization of word distribution across sensory, emotion, and association aspects.

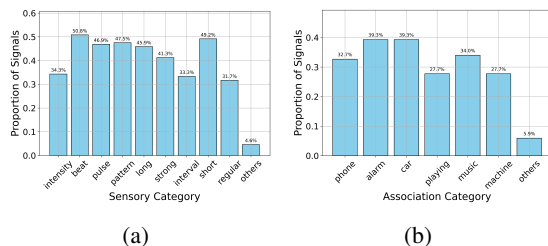


Figure 8: Distribution of vibration signals in HapticCap over (a) sensory words, and (b) associative words.

	Distinct-1/1-Gram $\uparrow$	Distinct-2/2-Gram $\uparrow$	Distinct-3/3-Gram $\uparrow$
Sensory	0.0599/7.7289	0.3528/12.0394	0.6793/13.9565
Emotion	0.0639/6.9261	0.2995/10.2775	0.5303/11.8154
Association	0.0868/7.9509	0.4065/11.8475	0.6794/13.5471

Table 9: Metrics on distinct-n and n-gram of description.

C Data Collection and Compensation

In this work, the annotators come from diverse backgrounds, including students, haptic designers,

Rank	Retrieved Description
1	The sensation is a sporadic sharp vibration and it is very alarming.
2	The sensation is like subtle buzzes.
3	The vibrations are strong with random stronger vibrations.
4	This sensation is light, irregular and lumpy.
5	The sensation is long repeating vibration lines.
6	This sensation is fuzzy and tender.
7	This sensation is smooth and soft.
8	This sensation is discontinued.
9	The sensation is very light vibration.
10	The sensation was regular and low intensity.

Table 10: Retrieved descriptions by rank.

and researchers. Figure 9 visualizes participant backgrounds. Each participant described 16 signals (one group) within one hour and received \$15 USD in cash or an equivalent value in gifts (e.g., chocolate) as compensation for their time. This compensation rate is above the minimum local wage. The data collection period spans approximately 11 months, including the user study and the creation of new haptic signals.

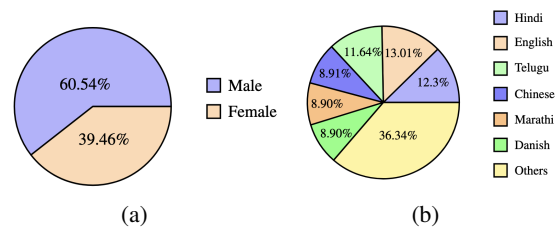


Figure 9: Distribution of (a) gender and (b) native language of users collected in the HapticCap dataset. All users were fluent English speakers.

To ensure high-quality annotations, we collected this dataset by having users feel vibrations on VR handheld devices in a controlled lab setting. Figure 10 shows the graphical user interface that participants used to play the haptic signals and includes the questions and instructions to write sensory, emotional, and associative descriptions. Table 11 presents example sensory, emotional, and associative descriptions from three different participants in our dataset.

We divide all haptic signals into 19 groups/sets, with each set containing 16 signals. The first two sets have a larger number of participants as we collected more descriptions for these two sets of haptic signals when we started the data collection. Based on our initial analysis of these descriptions, we found that descriptions from 10 users per signal provide enough variation, and thus we continued with describing each signal by 10 users for the fol-

lowing sets. We include the additional descriptions for the first two sets in HapticCap to enable future work to do further analysis of the individual differences and subjectivity in haptic descriptions.

	Category	Description
P_1	sensory	Its frequency is not very strong but very regular.
	emotion	I feel little bit impatient.
	association	It reminds me of my heartbeat after an intense workout.
P_2	sensory	feeling is like beating of a heart that is beating fast.
	emotion	makes me feel a bit stressed , like I need to hurry.
	association	It is like the heartbeat after running fast.
P_3	sensory	The sensation was short and with a light vibration.
	emotion	I feel being made hurrying up by the vibrations.
	association	It reminds me of the feeling of sitting on a motor boat.

Table 11: Example haptic descriptions in sensory, emotional, and associative categories. P_i denotes participant ID.

Figure 10: Illustration of the data collection interface.