# MASKTAC: MASK-GUIDED LEARNED LOSSY COMPRESSION FOR SPARSE TACTILE GLOVE DATA

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Tactile sensing provides direct feedback about physical contact for a wide range of perception and interaction applications. However, high-quality force-based tactile datasets remain limited. The widespread use of tactile data also creates a growing need of efficient compression for storage and transmission. To mitagate these gaps, we firstly introduce TacGlove, a tactile glove dataset collected using a customized wearable tactile glove. It contains 4,660 sequences and 781,969 frames, covering 32 object classes and 8 grasp types. The glove records three-axis contact forces from 1,207 taxels distributed across the hand. Such force-based tactile glove data exhibit pronounced spatiotemporal sparsity and strong correlations within contact regions. Based on these characteristics, we then propose MaskTac, a learned lossy codec for sparse tactile glove data. By combining sparsity-aware quantization with a mask-guided contextual entropy model, MaskTac exploits data spatiotemporal sparsity and reduces spatiotemporal redundancy within contact regions for efficient compression. We further establish a benchmark for tactile compression by adapting traditional and learned visual codecs to TacGlove. The benchmark evaluates both reconstruction fidelity and object and grasp classification accuracy. Experiments show that MaskTac outperforms the benchmarked codecs, achieving BD-rate improvement relative to VTM of 34.04% under PSNR, 32.57% under MS-SSIM, and 58.83% under Contact F1. Ablation studies further validate the effectiveness of the proposed design.

## 1 INTRODUCTION

Tactile sensing is fundamental to embodied intelligence, providing direct feedback about physical contact and interaction dynamics. It supports stable grasping, slip detection, and dexterous manipulation, including in-hand object rotation and the handling of thin, deformable objects (Yin et al., 2023; Lin et al., 2025). Research on these tasks requires datasets that capture real-world contact interactions. Existing tactile datasets span different sensing modalities. For example, Touch and Go uses optical images to record contact-induced deformation of the sensor surface (Yang et al., 2022). Dex-MPTD records distributed contact forces using a robotic hand (Zhao et al., 2025). Wearable tactile gloves can directly capture natural human interactions across diverse objects and grasp configurations. STAG records whole-hand contact distributions using 548 sensing elements, but primarily measures normal forces (Sundaram et al., 2019). OSMO supports three-axis tactile sensing, but covers only 12 sensing locations (Yin et al., 2025). Therefore, force-based tactile datasets combining dense spatial coverage, continuous three-axis force measurements, and diverse human grasps remain limited.

As tactile sensing becomes more widely used, the demand for storing and transmitting tactile data continues to grow. Dense tactile sensor arrays require communication links to transfer measurements from local acquisition units to central processors (Cannata et al., 2008; Giovinazzo et al., 2024). Beyond communication within sensing systems, teleoperation and immersive interaction require tactile information to be transmitted across devices to provide remote contact feedback and interactive haptic experiences (Luo et al., 2024; Yu et al., 2019; IEEE, 2024). Large-scale collection of human demonstrations also incurs substantial storage costs. For illustration, suppose each hand is equipped with 1,207 three-axis taxels sampled at 30 Hz, with each axis stored as a 32-bit value. In a collection campaign involving 1,000 participants, recording both hands for eight hours per day over 60 days would generate approximately 1.5 PB of uncompressed tactile measurements alone. Such costs

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

accumulate with the scale and duration of data collection, highlighting the need for efficient tactile compression.

Force-based tactile glove data exhibit pronounced spatial and temporal sparsity. Spatially, contact is confined to a subset of the taxels distributed across the fingers and palm. Temporally, individual taxels are active only during parts of an interaction, leaving many locations and time intervals without contact. During sustained contact, neighboring taxels often exhibit correlated force responses, while contact patterns and force values remain correlated across successive frames. The sparse distribution of contact and the spatiotemporal redundancy within contact regions together offer substantial compression potential. Existing tactile compression studies have largely explored adapting visual codecs to tactile data. For optical tactile data, these approaches apply existing visual codecs or retrain them on tactile images. For force-based data, TacCompress and TaCo organize multi-point force measurements into images and evaluate existing codecs (Zhao et al., 2025; Cheng et al., 2026). Although visual coding approaches achieve strong performance on optical tactile data, the advantages of learned visual codecs do not consistently extend to force measurements. For example, TacCompress reports that pretrained TCM underperforms conventional codecs HM and VTM (Zhao et al., 2025). These findings suggest that directly transferring visual codecs is insufficient to fully exploit the compression potential of sparse force-based tactile data. This motivates the development of a dedicated learned lossy codec that exploits tactile sparsity while reducing spatiotemporal redundancy within contact regions.

To address these limitations, we introduce **TacGlove**, a tactile glove dataset comprising 4,660 grasping sequences and 781,969 frames of three-axis force measurements, covering 32 object classes and 8 grasp types. The dataset is collected using a customized wearable tactile glove equipped with 1,207 force sensors (taxels), organized into arrays that follow the spatial layout of the hand. During collection, the wearer performs object grasping while the glove continuously records the spatial distribution and temporal evolution of contact forces. To efficiently compress the sparse tactile glove data, we propose **MaskTac**, a learned lossy codec for sparse tactile glove data. By combining sparsity-aware quantization with a mask-guided contextual entropy model, MaskTac exploits tactile sparsity and spatiotemporal dependencies within contact regions to achieve efficient compression while preserving contact information.

To evaluate compression performance on sparse tactile glove data, we establish a tactile compression benchmark by adapting traditional and learned visual codecs to TacGlove. The benchmark evaluates tactile reconstruction fidelity using PSNR, MS-SSIM, and Contact F1. We further assess task performance using object and grasp classification models trained on uncompressed tactile data and kept frozen when evaluating reconstructions from different codecs. Experiments demonstrate that MaskTac achieves state-of-the-art compression performance, with BD-rate of 34.04%, 32.57%, and 58.83% aginst VTM under PSNR, MS-SSIM, and Contact F1, respectively.

Generally speaking, our contributions can be summarized as

- **A tactile glove dataset.** We construct TacGlove by recording human grasping with a customized wearable tactile glove. The dataset contains 4,660 sequences comprising 781,969 frames of three-axis tactile maps, annotated with labels for 32 object classes and 8 grasp types.
- **A learned lossy codec tailored for sparse tactile glove data.** We propose MaskTac, which combines sparsity-aware quantization with a mask-guided contextual entropy model to achieve efficient tactile compression by exploiting sparsity and spatiotemporal redundancy in tactile signals. Extensive ablation studies are conducted to validate their effectiveness.
- **A tactile compression benchmark.** We adapt traditional and learned visual codecs to establish a benchmark that evaluates tactile compression performance from reconstruction fidelity and accuracy on object and grasp classification tasks. Experiments shows our proposed MaskTac achieves state-of-the-art compression performance.

## 2 Related Work

**Tactile Datasets.** Optical tactile datasets record sensor deformation as images. Touch and Go and SSVTP pair touch with vision (Yang et al., 2022; Kerr et al., 2023). TVL adds language annotations (Fu et al., 2024). ObjectFolder 2.0 provides simulated multisensory observations (Gao

et al., 2022). More recently, ToucHD expands optical tactile collection to dynamic interactions (Feng et al., 2026). Force-based datasets record contact measurements directly. Dex-MPTD captures three-axis forces at 1,140 taxels during robotic-hand grasping (Zhao et al., 2025), while ObjTac records local object contacts with a 60-taxel three-axis sensor (Cheng et al., 2025; 2026). Wearable gloves complement robot-collected data by directly recording contact patterns and force dynamics shaped by human hand movements and grasp strategies. STAG records normal-force distributions at 548 taxels during object grasping (Sundaram et al., 2019). OpenTouch measures normal pressure at 169 finger and palm taxels during everyday grasping and manipulation (Song et al., 2025). EgoTac-SC records normal-pressure signals at 132 locations per hand (264 in total) during tasks such as folding clothes and wiping tables (Zhang et al., 2026b). ART-Glove measures normal contact magnitudes at 2,048 taxels during ball rotation, bottle-cap screwing, and pressing (Lin & Zhao, 2026). OSMO supports normal and shear force measurements at 12 fingertip and palm locations, with its wiping demonstrations using only the five fingertip sensors (Yin et al., 2025). In this work, we introduce TacGlove, a tactile glove dataset comprising continuous human grasp sequences with three-axis measurements at 1,207 dense taxels and labels for 32 object classes and 8 grasp types.

**Tactile Data Compression.** For optical tactile images, GelSight compression separates lossy shape-texture features from losslessly coded force-marker information (Zeng et al., 2026), while OmniZip learns a multimodal lossless compressor evaluated on Touch and Go and ObjectFolder (Zhao et al., 2026). For force measurements, TacCompress benchmarks image codecs on taxel-time representations of robotic-hand data (Zhao et al., 2025). TaCo evaluates both optical and force-based tactile data. It introduces TaCo-LL, a lossless codec adapted from DualComp-I, and TaCo-L, a lossy codec adapted from LALIC. Both are retrained on optical tactile datasets (Cheng et al., 2026). These adaptations achieve strong compression on optical tactile images. However, the compression performance of both pretrained and fine-tuned learned lossy image codecs remains limited on sparse force data, as observed in prior benchmarks (Zhao et al., 2025; Cheng et al., 2026) and our experiments. Dedicated learned lossy compression for sparse force-based tactile data remains underexplored. To address this gap, we design MaskTac, a learned lossy codec tailored to the sparsity and spatiotemporal correlations of tactile glove data. It achieves state-of-the-art compression performance on the proposed tactile compression benchmark built on TacGlove.

# 3 TacGlove: Tactile Glove Dataset

## 3.1 Acquisition Setup

**Acquisition Device.** As shown in Figure 1(a), we use a customized wearable tactile glove to collect the TacGlove dataset. Hall-effect-based tactile sensors distributed across the hand capture contact signals as the wearer grasps and manipulates objects. We record these signals at 30 Hz from 1,207 sensing locations (taxels). Each taxel records the X, Y, and Z components of the contact force during manipulation. The X and Y components measure tangential forces, while the Z component measures the normal force in local coordinate system of each sensor. The taxels are organized into 14 two-dimensional arrays of different sizes to fit different regions of the hand. Detailed array configurations are provided in Appendix A.1.

**Collection Protocol.** During data collection, participants wear the tactile glove and grasp objects using predefined grasp types. Tactile signals are recorded continuously throughout each interaction. Before collection, participants receive systematic training and practice. TacGlove covers 8 grasp types and 32 object classes, as shown in Figure 1(b) and (c). The objects include containers, balls, fruit, packaged goods, fabric, and utensils. Variations in object shape, material, and grasp type produce different contact patterns across the hand. Each sequence is labeled with the category of the grasped object and the grasp type. Appendix A.2 provides the complete label catalog.

## 3.2 Dataset Composition and Split

**Data Representation.** TacGlove contains 4,660 tactile sequences comprising 781,969 frames, with sequence lengths ranging from 70 to 544 frames. At 30 Hz, consecutive frames are approximately 33.3 ms apart. Each tactile sequence contains measurements from 14 tactile sensor arrays. We represent a tactile sequence of $T$ frames as $\mathcal{X}$, with $X_r$ denoting the measurements from array $r$:

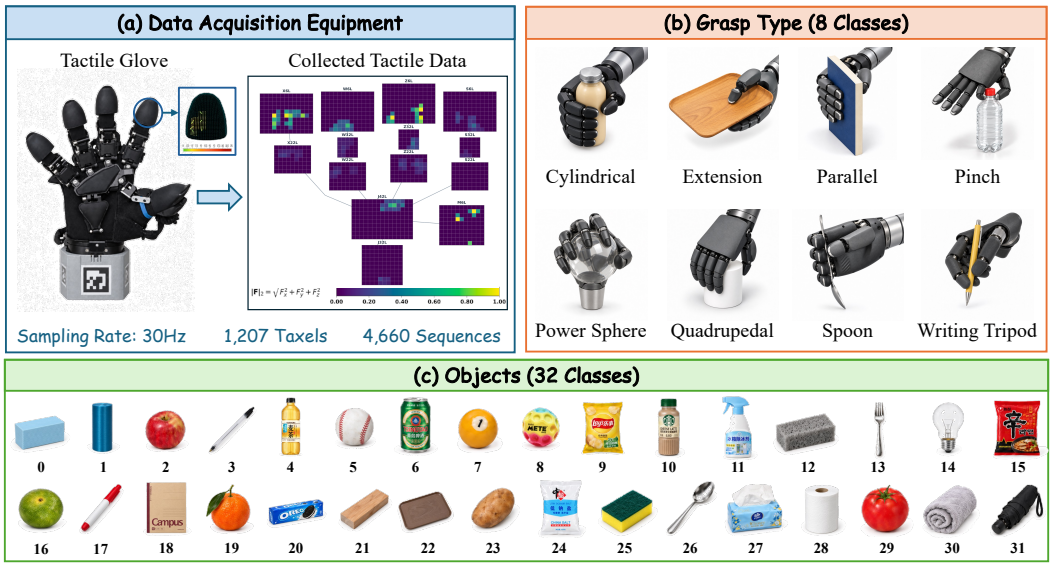


Figure 1: TacGlove Acquisition Overview. **(a)** Acquisition device and tactile measurements. **(b)** 8 grasp types. **(c)** 32 object categories.

$$\mathcal{X} = \{X_r\}_{r=1}^{14}, \quad X_r \in \mathbb{R}^{T \times H_r \times W_r \times 3}, \quad \sum_{r=1}^{14} H_r W_r = 1{,}207 \tag{1}$$

where $r$ indexes the tactile sensor arrays, and $H_r$ and $W_r$ denote the height and width of array $r$. The final dimension stores the X, Y, and Z components of contact force at each taxel.

**Dataset Split.** We randomly split the dataset into training and test sets at an 8:2 ratio with 3,728 and 932 sequences. Each sequence is kept intact within one set to prevent leakage between training and test sets through frames from the same sequence. Both sets cover all 32 object classes and 8 grasp types. Category labels and per-category split counts are provided in Appendix A.2.

### 3.3 Data Characteristics

**Temporal Sparsity.** Contact occurs intermittently as the hand approaches, grasps, and releases an object. Before contact and after release, taxels are typically inactive (all three recorded XYZ force components are zero). A continuously recorded tactile sequence therefore contains intervals with and without tactile responses. The left panel of Figure 2 shows how the number of nonzero taxels rises and falls during interaction and returns to zero when contact ends. Within contact intervals, responses can persist across consecutive frames, creating temporal redundancy alongside sparsity.

**Spatial Sparsity.** Tactile signals are also localized in space. Object shape and grasp posture determine which parts of the fingers and palm are engaged, leaving much of the sensing surface inactive even during a grasp. This produces spatially sparse measurements, with responses concentrated in a small subset of taxels. The right panel of Figure 2 illustrates this pattern: contact responses occupy localized bands, while most taxel positions remain zero.

Spatiotemporal sparsity is an intrinsic property of tactile glove data. Inactive taxel positions and intervals are part of the tactile representation, describing where and when tactile responses are absent. Tactile glove data also complement visual observations in downstream embodied tasks, including vision-tactile-language-action learning and multimodal demonstration learning (Zhang et al., 2025; Cheng et al., 2025; Fan et al., 2026). Removing inactive positions or intervals without preserving their original indices can disrupt alignment between these modalities. At the same time, storing and transmitting complete tactile sequences entails repeatedly representing large numbers of inactive measurements. This motivates dedicated tactile compression methods that exploit spatiotemporal sparsity and redundancy to reduce storage and transmission costs while preserving contact information and the structure needed for multimodal alignment.

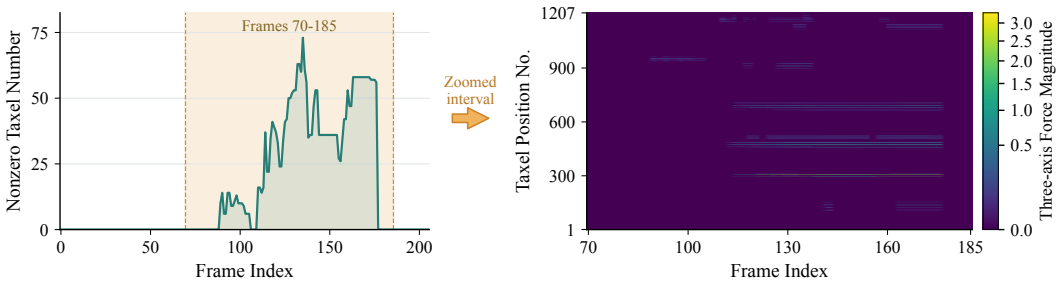


Figure 2: Spatiotemporal Sparsity of Tactile Glove Data. **Left:** the number of nonzero taxels over a 206-frame sequence recorded during a cylindrical grasp of a coffee bottle. Orange shading highlights frames 70-185. **Right:** the tactile distribution within this interval, showing the three-axis force magnitude at all 1,207 taxel positions.

# 4 MaskTac: Learnable Lossy Tactile Codec

## 4.1 Overview

Figure 3 presents the MaskTac framework for encoding and decoding tactile glove data. For a sequence $\mathcal{X}$ of $T$ tactile frames, each frame $x_t \in \mathbb{R}^{H\times W\times 3}$ is represented as a tactile map by padding and arranging the native sensor arrays according to the sensor layout on the hand, where $H$ and $W$ denote the map height and width. The three channels represent the amplitudes along the X, Y, and Z axes. At the encoding end, the amplitudes first undergo sparsity-aware quantization controlled by the quantization parameter $k$. The mask $m_t$ then identifies nonzero taxels, whose amplitudes are encoded using the proposed mask-guided contextual entropy model. The side information required by quantization is included in the header of the bitstream, and the binary mask is encoded losslessly. AE and AD denote arithmetic encoding and decoding (Rissanen, 1976; Rissanen & Langdon, 1981). At the decoding end, the model reads the side information and decodes the mask and amplitudes in turn. Inverse quantization then yields the reconstructed tactile frame $\hat{x}_t$. Taxels excluded by the decoded mask are reconstructed as zeros, while decoded amplitudes are restored to their original positions.

## 4.2 Sparsity-Aware Quantization

To preserve the sparse structure of tactile data, we use uniform quantization that preserves zero values exactly. The entire sequence shares the step size $\Delta_k = r2^{-k}$. Here $r = x_{\max} - x_{\min}$ is the numerical range. The values $x_{\max}$ and $x_{\min}$ denote the maximum and minimum amplitudes over the entire sequence, respectively. Quantization and inverse quantization are

$$\bar{x}_t = \mathrm{round}(x_t/\Delta_k), \qquad \hat{x}_t = \Delta_k \bar{x}_t \tag{2}$$

This quantizer preserves the original zero values while discretizing nonzero response amplitudes, maintaining sparsity for subsequent compression. The quantization parameter $k$ controls the rate-distortion trade-off through the step size. Larger $k$ values reduce quantization error and improve reconstruction quality, typically at the cost of a higher bitrate. The quantization side information consists of $k$, $x_{\min}$, and $x_{\max}$ and is stored in the bitstream header. The decoder uses these values to recover $\Delta_k$ for inverse quantization.

## 4.3 Mask-Guided Contextual Entropy Model

**Mask Extraction and Coding.** For each taxel $n$ in the quantized frame $\bar{x}_t$, we extract a binary mask as

$$m_t(n) = \begin{cases} 1, & \text{if } \max_{c\in\{X,Y,Z\}} |\bar{x}_t(n,c)| > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

where $\bar{x}_t(n,c)$ denotes the quantized amplitude along axis $c \in \{X,Y,Z\}$ at taxel $n$ in frame $t$. We encode the mask $m_t \in \{0,1\}^{H\times W}$ at frame $t$ losslessly with a binary entropy coder. The mask selects taxels for amplitude coding and serves as part of the context for predicting amplitude distributions.

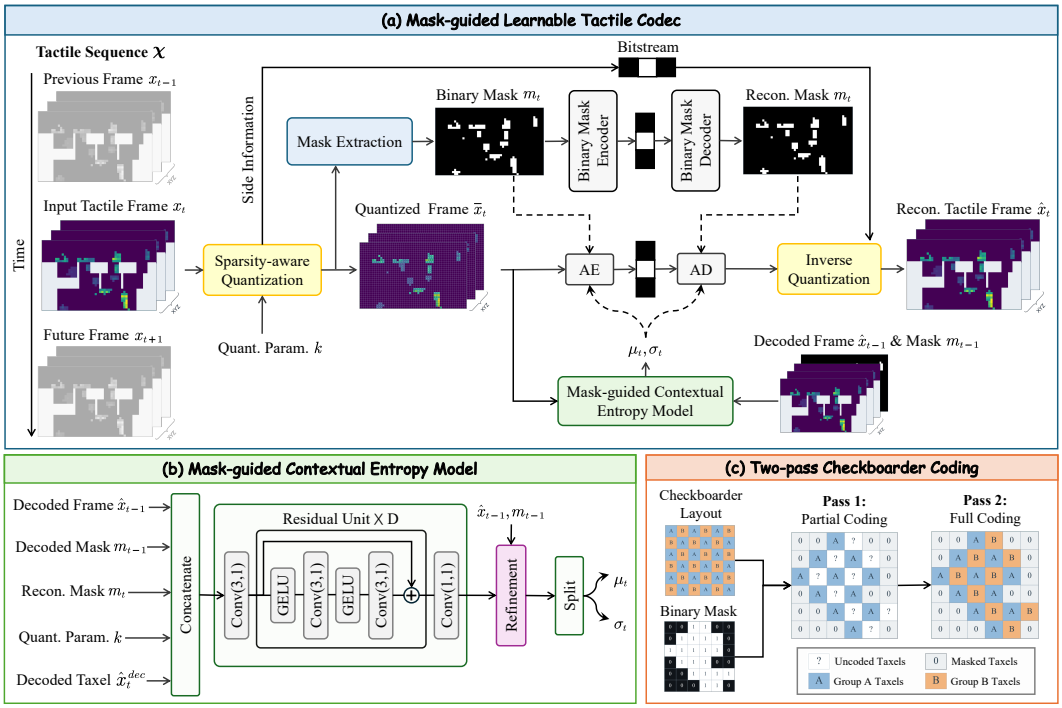


Figure 3: Illustration of the MaskTac Framework. **(a)** Encoding and decoding pipeline. **(b)** Architecture of the mask-guided contextual entropy model. **(c)** Two-pass checkerboard coding mechanism.

**Contextual Entropy Model.** As shown in Figure 3(b), the proposed mask-guided contextual entropy model uses context information to estimate the mean $\mu_t$ and scale $\sigma_t$ of a Gaussian distribution $\mathcal{N}(\mu_t, \sigma_t^2)$ for each XYZ component. Discretizing these distributions yields symbol probabilities for arithmetic coding of the quantized amplitudes. Probability prediction covers the full tactile map, while amplitude coding is restricted to mask-selected nonzero taxels. The prediction network consists of a $3 \times 3$ input convolution, $D$ stacked residual blocks, and a $1 \times 1$ output convolution, where $D$ controls the network depth. The context $C_t$ at frame $t$ are concatenated as

$$C_t = \mathrm{Concat}\left[\hat{X}_t^{\mathrm{hist}}, M_t^{\mathrm{hist}}, m_t, \hat{x}_t^{\mathrm{dec}}, \mathbf{k}\right] \tag{4}$$

where $\hat{X}_t^{\mathrm{hist}}$ stacks the previous $L$ reconstructed frames and $M_t^{\mathrm{hist}}$ stacks their masks. The history length $L$ controls the number of past frames used by the model. The constant map $\mathbf{k} \in \mathbb{R}^{H \times W}$ is filled with the quantization parameter $k$. The tensor $\hat{x}_t^{\mathrm{dec}} \in \mathbb{R}^{H \times W \times 3}$ stores the already decoded amplitudes of the current frame, with undecoded positions set to zero. Missing historical frames and masks are zero-filled.

**Refinement.** We adopt a refinement strategy to estimate the distribution parameters. Specifically, the network in entropy model predicts a mean correction $\delta_t$ and a scale correction $\ell_t$ from the context rather than directly predicting the final mean $\mu_t$ and scale $\sigma_t$. Combined with the reconstructed amplitudes $\hat{x}_{t-1}$ and mask $m_{t-1}$ of previous frame, these corrections yield

$$\mu_t = \mathrm{round}(\hat{x}_{t-1}/\Delta_k + \delta_t)\,, \qquad \sigma_t = \mathrm{clip}\big(g_k(m_{t-1})e^{\ell_t}\big) \tag{5}$$

Here, $g_k$ denotes the mask-based scale prior, and $\mathrm{clip}(\cdot)$ bounds the scale for numerical stability.

**Two-Pass Checkerboard Coding.** As shown in Figure 3(c), we partition the taxels in each tactile map into checkerboard groups $A$ and $B$, following prior checkerboard context models (He et al., 2021; 2022). In the first pass, group $A$ is coded with $\hat{x}_t^{\mathrm{dec}} = 0$. Its reconstructed amplitudes then populate $\hat{x}_t^{\mathrm{dec}}$, providing spatial context for coding group $B$ in the second pass. Both passes share the same network, and decoding follows the same order.

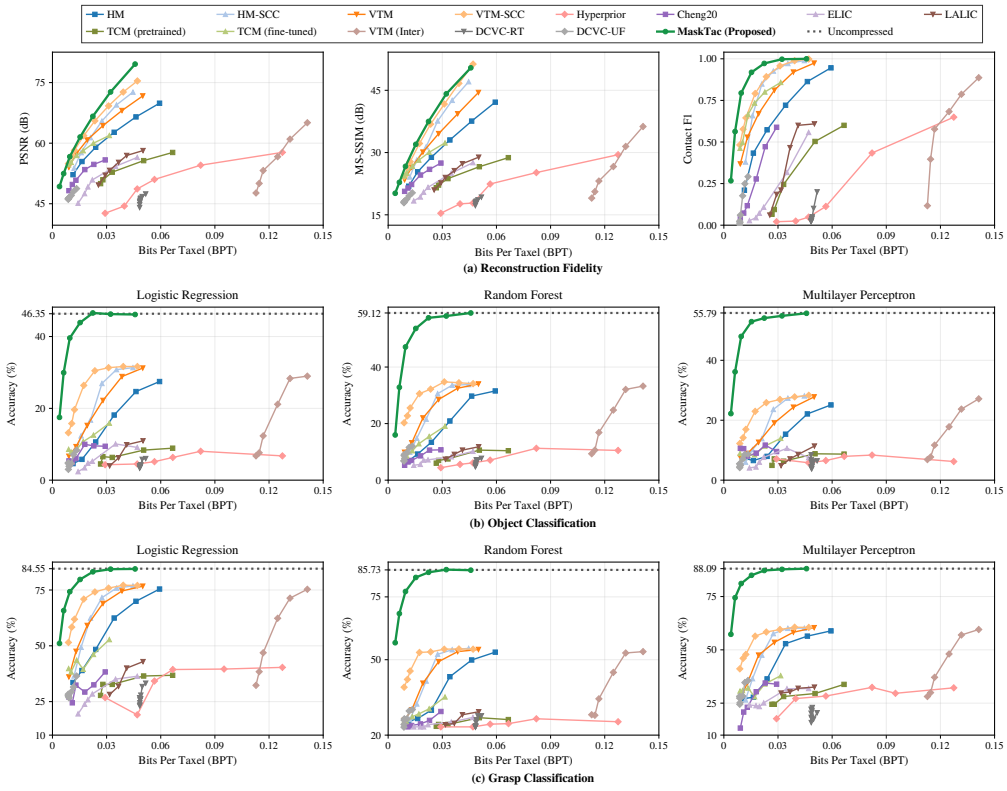


Figure 4: Rate-distortion Curves of Tactile Compression Performance on TacGlove dataset. **(a)** Reconstruction Fidelity. **(b)** Object Classification. **(c)** Grasp Classification. Dotted horizontal lines indicate classification accuracy on uncompressed tactile data.

### 4.4 Training

**Loss Function.** We train the entropy model to reduce the bits required to encode amplitudes at nonzero taxels. To this end, we minimize the negative log-likelihood of their quantized amplitudes:

$$\mathcal{L} = -\frac{1}{N} \sum_{i \in \Omega_t} \log_2 p_\theta(\bar{x}_{t,i} \mid C_{t,i}) \tag{6}$$

where $\Omega_t$ indexes all XYZ components of mask-selected nonzero taxels, $N$ is the total number of native taxels per frame. The $C_{t,i}$ is the context available when encoding component $i$.

**Implementation Details** We train the model on the TacGlove training set using Adam for 100 epochs with a batch size of 32. The initial learning rate is $10^{-3}$ and is halved at the start of epochs 61 and 86. We use $L = 2$ historical frames and $D = 1$ residual block to balance compression performance and model size. Appendix C presents ablation studies supporting these choices. We set $k \in \{1, 2, 3, 4, 5, 6, 7\}$ to obtain multiple bitrate operating points. The same trained model parameters are shared across all values of $k$.

## 5 Experiments

### 5.1 Benchmark for Tactile Compression

To evaluate the tactile compression performance of MaskTac, we adapt traditional and learned image and video codecs to tactile glove data, constructing a benchmark comprising ten image codec and three video codec configurations.

Table 1: BD-rate (%) of Tactile Compression Performance on TacGlove dataset. **Take VTM as the anchor.** CF1 denotes Contact F1. TCM(PT) and TCM(FT) denote pretrained and fine-tuned TCM.

| Metric | VTM | HM | HM-SCC | VTM-SCC | Cheng20 | TCM(PT) | TCM(FT) | VTM(Inter) | MaskTac |
|---|---|---|---|---|---|---|---|---|---|
| PSNR | 0.00 | 42.92 | -6.02 | -23.71 | 144.86 | 370.37 | 28.35 | 627.19 | **-34.04** |
| MS-SSIM | 0.00 | 37.66 | -10.59 | -25.12 | 80.65 | 272.89 | 9.00 | 613.70 | **-32.57** |
| CF1 | 0.00 | 54.14 | -13.42 | -31.62 | 102.42 | 318.31 | -20.67 | 612.74 | **-58.83** |

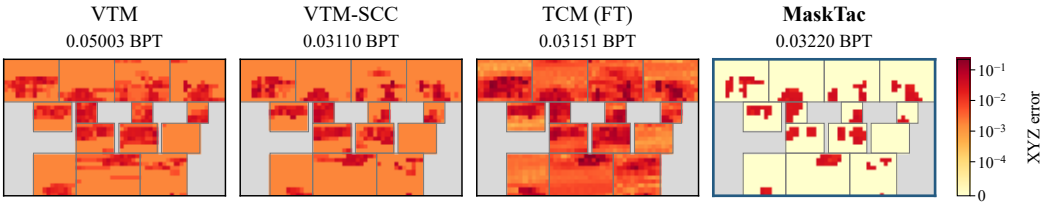


Figure 5: Reconstruction Errors for the Tactile Frame during a Power Sphere Grasp of Salt Bag. Gray regions are padding. MaskTac preserves the zero-valued background exactly.

**Metrics.** We evaluate reconstruction quality of tactile data using PSNR and MS-SSIM (Wang et al., 2003). We also report Contact F1 (Zhang et al., 2026a) to measure agreement between reference and reconstructed tactile contact maps. We measure bitrate in Bits Per Taxel (BPT), following the use of Bits Per Pixel (BPP) in visual compression. BPT is the average number of compressed bits per taxel per frame. Metric details are provided in Appendix B.1.

**Classification.** We evaluate object and grasp recognition using logistic regression (LR) (Hastie et al., 2009), random forest (RF) (Breiman, 2001), and a multilayer perceptron (MLP) (Goodfellow et al., 2016). These classifiers are trained on the uncompressed training sequences of TacGlove and then frozen to measure classification accuracy after compression. Details are given in Appendix B.2.

**Baselines.** Image compression baselines compress each tactile sequence by arranging it into a taxel-time image $I \in \mathbb{R}^{1207\times T\times 3}$, as shown in Figure 2 (Right). These baselines include HM, HM-SCC (JVET, 2026a), VTM, VTM-SCC (JVET, 2026b), Hyperprior (Ballé et al., 2018), Cheng20 (Cheng et al., 2020), ELIC (He et al., 2022), TCM (Liu et al., 2023) and LALIC (Feng et al., 2025). We evaluate both pretrained TCM and a version fine-tuned on the TacGlove training set for 200 epochs. Video compression baselines use the same representation as MaskTac, as shown in Figure 3(a). The native sensor arrays are arranged and padded to form the hand-shaped tactile map sequence $S \in \mathbb{R}^{T\times 32\times 52\times 3}$. These baselines include VTM (Inter) (JVET, 2026b), DCVC-RT (Jia et al., 2025), and DCVC-UF (Li et al., 2026). Codec configurations are detailed in Appendix B.3.

### 5.2 Compression Performance

**Reconstruction Fidelity.** Figure 4(a) compares tactile compression performance on the TacGlove test set. Table 1 reports BD-rate (Bjøntegaard, 2001) relative to VTM. MaskTac achieves the best performance across all three metrics, reducing bitrate by **34.04%** under PSNR, **32.57%** under MS-SSIM, and **58.83%** under Contact F1. VTM with screen content coding tools (VTM-SCC) ranks second to MaskTac across all three metrics, reducing bitrate by 23.71% under PSNR, 25.12% under MS-SSIM, and 31.62% under Contact F1. In contrast, directly applying learned visual codecs such as Cheng20 and DCVC-UF yields poor rate-distortion performance. Fine-tuning TCM on TacGlove substantially improves its performance, achieving a 20.67% bitrate reduction under Contact F1. However, its BD-rate remains 28.35% above VTM under PSNR and 9.00% above VTM under MS-SSIM. These results suggest that tactile data have distinct characteristics that require tailored codec designs for efficient compression. VTM (Inter) also requires substantially higher bitrate, indicating that temporal coding alone does not guarantee efficient tactile compression.

**Classification Performance.** Figure 4(b) and (c) report object and grasp classification accuracy after tactile data compression. Across the three classifiers, MaskTac achieves a substantially better

rate-accuracy trade-off than the visual baselines on both tasks. At the higher tested bitrates, its classification accuracy approaches that of uncompressed tactile data.

Table 2: Ablation Study on MaskTac. **Take Full MaskTac as the anchor of BD-rate(%).** LR, RF, and MLP denote logistic regression, random forest, and multilayer perceptron.

| Model Variant | Recon. Fidelity | | | Object Classification | | | Grasp Classification | | |
|---|---|---|---|---|---|---|---|---|---|
| | PSNR | MS-SSIM | CF1 | LR | RF | MLP | LR | RF | MLP |
| Full MaskTac (anchor) | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *Entropy Model Context* | | | | | | | | | |
| w/o history & decoded taxels | 192.89 | 194.23 | 199.45 | 200.70 | 197.64 | 191.01 | 201.11 | 198.76 | 199.38 |
| w/o mask conditioning | 3.99 | 4.02 | 4.70 | 4.72 | 4.59 | 4.86 | 4.60 | 4.77 | 4.62 |
| w/o quant. param. $k$ | 6.12 | 6.12 | 9.65 | 9.92 | 9.39 | 10.10 | 8.98 | 10.08 | 9.31 |
| *Coding Mechanisms* | | | | | | | | | |
| w/o mask guidance | 102.18 | 100.92 | 78.58 | 76.75 | 81.60 | 80.47 | 80.63 | 77.10 | 80.52 |
| w/o refinement | 9.44 | 9.05 | 4.54 | 4.07 | 5.32 | 5.66 | 4.65 | 4.38 | 4.89 |
| w/o two-pass coding | 6.06 | 5.98 | 4.23 | 4.13 | 4.49 | 4.16 | 4.46 | 4.08 | 4.42 |

**Visualization.** Figure 5 visually compares reconstruction errors across methods. With its dedicated mask bitstream, MaskTac exactly preserves the zero-valued background and inactive taxels, while also achieving lower reconstruction errors in nonzero taxels.

### 5.3 Ablation Study

To evaluate the effectiveness of components in MaskTac, we conduct extensive ablation studies. Table 2 reports the changes in compression performance by BD-rate on the TacGlove test set.

**Entropy Model Context.** MaskTac uses historical frames and already decoded taxels to exploit temporal and within-frame correlations in tactile signals. Removing both sources of context increases BD-rate by 192.89%, 194.23%, and 199.45% under PSNR, MS-SSIM, and Contact F1, and by 191.01-201.11% across the classification metrics. This is the largest degradation among the evaluated variants, identifying contextual modeling as a major contributor to compression performance of MaskTac. We further examine the mask and quantization parameter inputs. Removing only the current and historical mask inputs, increases BD-rate by 3.99-4.86% across all metrics. Removing the concatenated map $\mathbf{k}$ increases BD-rate by 6.12% to 10.10%. These results confirm the effectiveness of both mask and quantization parameter conditioning.

**Coding Mechanisms.** Mask guidance exploits tactile sparsity by encoding nonzero taxel locations in a mask bitstream and coding amplitudes only at these locations. The mask also serves as a context input to the proposed entropy model. To evaluate this mechanism, we completely remove the mask bitstream and mask conditioning. This increases BD-rate by 102.18%, 100.92%, and 78.58% under PSNR, MS-SSIM, and Contact F1, and by 76.75-81.60% across the classification metrics. The much larger penalty than removing mask inputs alone highlights the importance of the full mask-guided coding mechanism. Appendix C.2 analyzes how bitrate savings from the mask mechanism offset the additional mask-stream overhead. We also assess refinement and two-pass coding. Replacing refinement with direct prediction of $\mu$ and $\sigma$ increases BD-rate by 4.54-9.44% for reconstruction and 4.07-5.66% for classification. Replacing two-pass coding with single-pass coding increases BD-rate by 4.23-6.06% and 4.08-4.49%. Both mechanisms provide further gains by refining distribution estimates and making decoded amplitudes available as context.

## 6 Conclusion

In this study, we present TacGlove, a tactile glove dataset collected using a customized wearable tactile glove. We propose MaskTac, a learned lossy codec with sparsity-aware quantization and a mask-guided contextual entropy model. Furthermore, we establish a benchmark for tactile compression that evaluates reconstruction fidelity and classification accuracy. MaskTac achieves substantial bitrate savings over VTM and improves classification performance at comparable bitrates. These experimental results demonstrate the effectiveness of the proposed tactile compression method.

### AI USE STATEMENT

In this work, we used generative AI tools to assist with language polishing, including obtaining manuscript feedback through the LLM feedback tool provided by ICLR. We did not use generative AI tools to develop the research ideas, implement the proposed method, or obtain the experimental results. We have reviewed all AI-assisted work and revised the manuscript accordingly. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

### REPRODUCIBILITY STATEMENT

The TacGlove dataset introduced in this work, together with the source code for the proposed learned lossy tactile compression method MaskTac, will be released publicly upon acceptance.

## REFERENCES

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018.

Gisle Bjøntegaard. Calculation of average PSNR differences between RD-curves. Technical Report VCEG-M33, ITU-T Video Coding Experts Group (VCEG), Austin, TX, USA, April 2001.

Leo Breiman. Random forests. *Machine Learning*, 45:5-32, 2001. doi: 10.1023/A:1010933404324.

Giorgio Cannata, Marco Maggiali, Giorgio Metta, and Giulio Sandini. An embedded artificial skin for humanoid robots. In *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2008. doi: 10.1109/MFI.2008.4648033.

Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7939-7948, 2020.

Zhengxue Cheng, Yiqian Zhang, Anni Tang, Keyu Wang, Wenkang Zhang, Haoyu Li, Hengdi Zhang, and Li Song. OmniVTLA: Vision-tactile-language-action models with semantic-aligned tactile sensing. *arXiv preprint arXiv:2508.08706*, 2025.

Zhengxue Cheng, Yan Zhao, Keyu Wang, Hengdi Zhang, and Li Song. TaCo: A benchmark for lossless and lossy codecs of heterogeneous tactile data. In *International Conference on Learning Representations*, 2026.

Jiacheng Fan, Zhiyue Zhao, Yiqian Zhang, Chao Chen, Peide Wang, Hengdi Zhang, and Zhengxue Cheng. RoboPaint: From human demonstration to any robot and any view. *arXiv preprint arXiv:2602.05325*, 2026.

Donghui Feng, Zhengxue Cheng, Shen Wang, Ronghua Wu, Hongwei Hu, Guo Lu, and Li Song. Linear attention modeling for learned image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7623-7632, 2025.

Ruoxuan Feng, Yuxuan Zhou, Siyu Mei, Dongzhan Zhou, Pengwei Wang, Shaowei Cui, Bin Fang, Guocai Yao, and Di Hu. AnyTouch 2: General optical tactile representation learning for dynamic tactile perception. In *International Conference on Learning Representations*, 2026.

Letian Fu, Gaurav Datta, Huang Huang, William Chung-Ho Panitch, Jaimyn Drake, Joseph Ortiz, Mustafa Mukadam, Mike Lambeta, Roberto Calandra, and Ken Goldberg. A touch, vision, and language dataset for multimodal alignment. In *International Conference on Machine Learning*, volume 235, pp. 14080-14101, 2024.

Ruohan Gao, Zilin Si, Yen-Yu Chang, Samuel Clarke, Jeannette Bohg, Li Fei-Fei, Wenzhen Yuan, and Jiajun Wu. ObjectFolder 2.0: A multisensory object dataset for sim2real transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Francesco Giovinazzo, Francesco Grella, Marco Sartore, Manuela Adami, Riccardo Galletti, and Giorgio Cannata. From CySkin to ProxySKIN: Design, implementation and testing of a multi-modal robotic skin for human-robot interaction. *Sensors*, 24(4):1334, 2024. doi: 10.3390/s24041334.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition, 2009. doi: 10.1007/978-0-387-84858-7.

Dailan He, Yaoyan Zheng, Baocheng Sun, Yan Wang, and Hongwei Qin. Checkerboard context model for efficient learned image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14771-14780, 2021.

Dailan He, Ziming Yang, Weikun Peng, Rui Ma, Hongwei Qin, and Yan Wang. ELIC: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5718-5727, 2022.

IEEE. IEEE standard for haptic codecs for the tactile internet. IEEE Std 1918.1.1-2024, IEEE, June 2024.

Zhaoyang Jia, Bin Li, Jiahao Li, Wenxuan Xie, Linfeng Qi, Houqiang Li, and Yan Lu. Towards practical real-time neural video compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

JVET. HM: HEVC reference software, 2026a. URL `https://vcgit.hhi.fraunhofer.de/jvet/HM`. Software repository, accessed September 2026.

JVET. VTM: VVC reference software, 2026b. URL `https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM`. Software repository, accessed September 2026.

Justin Kerr, Huang Huang, Albert Wilcox, Ryan Hoque, Jeffrey Ichnowski, Roberto Calandra, and Ken Goldberg. Self-supervised visuo-tactile pretraining to locate and follow garment features. In *Robotics: Science and Systems*, 2023.

Jiahao Li, Wenxuan Xie, Zhaoyang Jia, Bin Li, Zongyu Guo, Xiaoyi Zhang, and Yan Lu. Ultra-fast neural video compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Changyi Lin and Ding Zhao. ART-Glove: Articulated tactile glove for contact-grounded dexterous interaction capture. *arXiv preprint arXiv:2606.16370*, 2026.

Pei Lin, Yuzhe Huang, Jianpeng Ma, Chenxi Xiao, Wanlin Li, and Ziyuan Jiao. PP-Tac: Paper picking using omnidirectional tactile feedback in dexterous robotic hands. In *Proceedings of Robotics: Science and Systems*, Los Angeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.056.

Jinming Liu, Heming Sun, and Jiro Katto. Learned image compression with mixed Transformer-CNN architectures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14388-14397, 2023.

Yiyue Luo, Chao Liu, Young Joong Lee, Joseph DelPreto, Kui Wu, Michael Foshey, Daniela Rus, Tomás Palacios, Yunzhu Li, Antonio Torralba, and Wojciech Matusik. Adaptive tactile interaction transfer via digitally embroidered smart gloves. *Nature Communications*, 15:868, 2024. doi: 10.1038/s41467-024-45059-8.

J. J. Rissanen. Generalized Kraft inequality and arithmetic coding. *IBM Journal of Research and Development*, 20(3):198-203, 1976. doi: 10.1147/rd.203.0198.

Jorma Rissanen and Glen Langdon. Universal modeling and coding. *IEEE Transactions on Information Theory*, 27(1):12-23, 1981. doi: 10.1109/TIT.1981.1056282.

Yuxin Ray Song, Jinzhou Li, Rao Fu, Devin Murphy, Kaichen Zhou, Rishi Shiv, Yaqi Li, Haoyu Xiong, Crystal Elaine Owens, Yilun Du, Yiyue Luo, Xianyi Cheng, Antonio Torralba, Wojciech Matusik, and Paul Pu Liang. OpenTouch: Bringing full-hand touch to real-world interaction. *arXiv preprint arXiv:2512.16842*, 2025. doi: 10.48550/arXiv.2512.16842.

Subramanian Sundaram, Petr Kellnhofer, Yunzhu Li, Jun-Yan Zhu, Antonio Torralba, and Wojciech Matusik. Learning the signatures of the human grasp using a scalable tactile glove. *Nature*, 569: 698-702, 2019. doi: 10.1038/s41586-019-1234-z.

Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers*, volume 2, pp. 1398-1402, 2003. doi: 10.1109/ACSSC.2003.1292216.

Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and Go: Learning from human-collected vision and touch. In *Advances in Neural Information Processing Systems*, 2022.

Jessica Yin, Haozhi Qi, Youngsun Wi, Sayantan Kundu, Mike Lambeta, William Yang, Changhao Wang, Tingfan Wu, Jitendra Malik, and Tess Hellebrekers. OSMO: Open-source tactile glove for human-to-robot skill transfer. *arXiv preprint arXiv:2512.08920*, 2025.

Zhao-Heng Yin, Binghao Huang, Yuzhe Qin, Qifeng Chen, and Xiaolong Wang. Rotating without seeing: Towards in-hand dexterity through touch. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.036.

Xinge Yu, Zhaoqian Xie, Yang Yu, Jungyup Lee, Abraham Vazquez-Guardado, Haiwen Luan, Jasper Ruban, Xin Ning, Aadeel Akhtar, Dengfeng Li, Bowen Ji, Yiming Liu, Rujie Sun, Jingyue Cao, Qingze Huo, Yishan Zhong, ChanMi Lee, SeungYeop Kim, Philipp Gutruf, Changxing Zhang, Yeguang Xue, Qinglei Guo, Aditya Chempakasseril, Peilin Tian, Wei Lu, JiYoon Jeong, YongJoon Yu, Jesse Cornman, CheeSim Tan, BongHoon Kim, KunHyuk Lee, Xue Feng, Yonggang Huang, and John A. Rogers. Skin-integrated wireless haptic interfaces for virtual and augmented reality. *Nature*, 575(7783):473-479, 2019. doi: 10.1038/s41586-019-1687-0.

Yaofeng Zeng, Chengdong Lan, and Yifeng Xu. GelSight dual-modal tactile data compression for machines. *Digital Signal Processing*, pp. 105696, 2026. doi: 10.1016/j.dsp.2025.105696.

Chaofan Zhang, Peng Hao, Xiaoge Cao, Xiaoshuai Hao, Shaowei Cui, and Shuo Wang. VTLA: Vision-tactile-language-action model with preference learning for insertion manipulation. *arXiv preprint arXiv:2505.09577*, 2025.

Shiqi Zhang, Xin Zhang, Yedong Shen, Yao Li, Yuxuan Gao, Sha Zhang, Yuan Zhang, Kaixue Long, Jiajia Wu, Jia Pan, Jiajun Deng, and Yanyong Zhang. ReTouch: Empowering contact-rich dexterous manipulation with online-refined tactile prediction. *arXiv preprint arXiv:2608.01824*, 2026a.

Wenkang Zhang, Chengbo Yuan, Zicheng Zhang, Zhengxue Cheng, and Yang Gao. EgoTac: In-the-wild tactile prediction from egocentric vision. *arXiv preprint arXiv:2608.15060*, 2026b.

Yan Zhao, Yang Li, Zhengxue Cheng, Hengdi Zhang, and Li Song. TacCompress: A benchmark for multi-point tactile data compression in dexterous hand. *arXiv preprint arXiv:2505.16289*, 2025. Version 2.

Yan Zhao, Zhengxue Cheng, Junxuan Zhang, Dajiang Zhou, Qunshan Gu, Qi Wang, and Li Song. OmniZip: Learning a unified and lightweight lossless compressor for multi-modal data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

# A TacGlove Dataset Details

## A.1 Tactile Sensor Configuration

TacGlove is collected using a Paxini instrumented glove (Fan et al., 2026). The glove contains 14 tactile sensor arrays with a total of 1,207 taxels, each recording the X, Y, and Z components of contact force. The arrays have different shapes and sizes to fit different regions of the hand. Table 3 summarizes their dimensions and taxel counts. Figure 6 illustrates the sensor array layout, with each array labeled by its ID in Table 3.

Table 3: Spatial dimensions and taxel counts of the tactile sensor arrays.

| **Sensor IDs** | $H \times W$ | **Arrays** | **Taxels** |
|---|---|---|---|
| J32L | $10 \times 10$ | 1 | 100 |
| J42L | $10 \times 15$ | 1 | 150 |
| M6L | $10 \times 11$ | 1 | 110 |
| S6L, Z6L, W6L, X6L | $10 \times 13$ | 4 | 520 |
| S22L, Z22L, W22L, X22L | $7 \times 9$ | 4 | 252 |
| S32L, Z32L, W32L | $5 \times 5$ | 3 | 75 |
| **Total** | - | **14** | **1,207** |

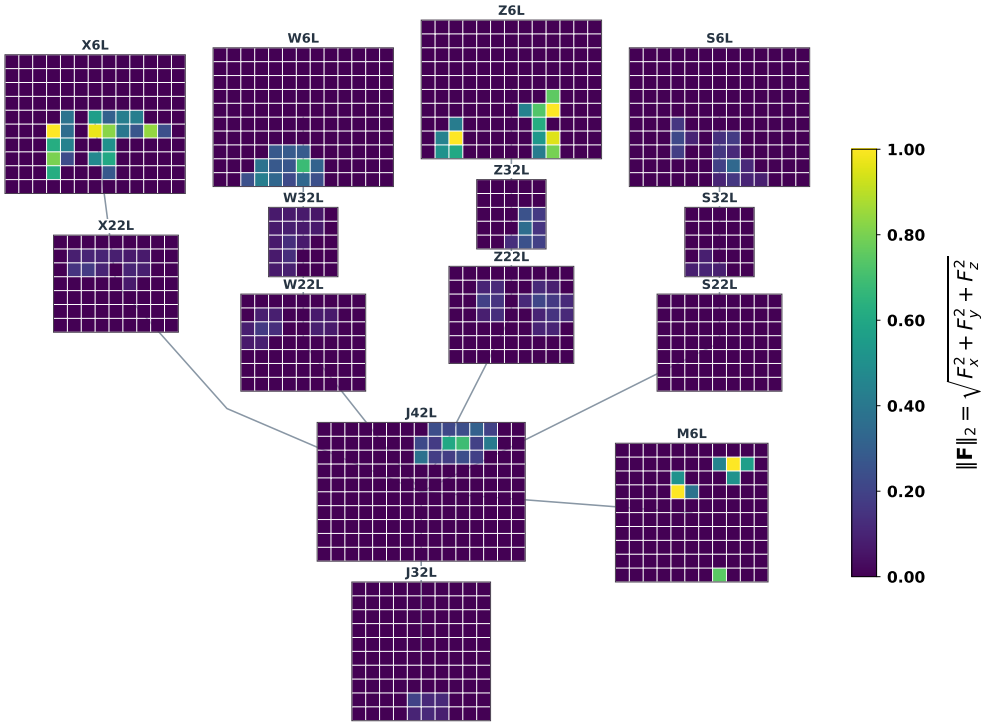


Figure 6: Tactile Sensor Array Layout And An Example Of The Contact Force Distribution (Power Sphere Grasp Of A Salt Bag).

## A.2 Dataset Composition

**Object and Grasp Categories.** TacGlove covers 32 object categories and 8 grasp types. Each sequence has one object label and one grasp-type label. Tables 4 and 5 list the category IDs and names, together with the number of training and test sequences in each category.

**Training and Test Splits.** The dataset contains 4,660 sequences comprising 781,969 frames. We split the dataset into training and test sets at an 8:2 ratio. The training set contains 3,728 sequences and 626,860 frames, while the test set contains 932 sequences and 155,109 frames. Each complete sequence belongs to only one split, and all 32 object categories and 8 grasp types appear in both sets. The number of sequences varies across object categories and grasp types, as detailed in Tables 4 and 5.

Table 4: Object categories and dataset split statistics for TacGlove.

| ID | Object Category | Train | Test | Total | ID | Object Category | Train | Test | Total |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 3D printed box | 135 | 41 | 176 | 16 | Mandarin | 115 | 32 | 147 |
| 1 | 3D printed cylinder | 140 | 35 | 175 | 17 | Marker | 17 | 3 | 20 |
| 2 | Apple | 113 | 33 | 146 | 18 | Notebook | 105 | 24 | 129 |
| 3 | Ballpen | 16 | 4 | 20 | 19 | Orange | 115 | 28 | 143 |
| 4 | Barley water | 395 | 110 | 505 | 20 | Oreo biscuit | 101 | 23 | 124 |
| 5 | Baseball | 84 | 22 | 106 | 21 | Plank | 101 | 24 | 125 |
| 6 | Beer bottle | 144 | 29 | 173 | 22 | Plastic plate | 64 | 16 | 80 |
| 7 | Billiards | 80 | 20 | 100 | 23 | Potato | 117 | 32 | 149 |
| 8 | Bouncy ball | 78 | 23 | 101 | 24 | Salt bag | 154 | 37 | 191 |
| 9 | Chips bag | 141 | 41 | 182 | 25 | Sponge | 156 | 34 | 190 |
| 10 | Coffee bottle | 142 | 33 | 175 | 26 | Spoon | 17 | 3 | 20 |
| 11 | Deicing agent | 145 | 34 | 179 | 27 | Tissue box | 144 | 29 | 173 |
| 12 | Foam cube | 152 | 36 | 188 | 28 | Tissue roll | 155 | 23 | 178 |
| 13 | Fork | 14 | 6 | 20 | 29 | Tomato | 117 | 30 | 147 |
| 14 | Glass bulb | 93 | 21 | 114 | 30 | Towel roll | 152 | 41 | 193 |
| 15 | Instant noodles bag | 139 | 44 | 183 | 31 | Umbrella | 87 | 21 | 108 |
| | | | | | | **Total (all 32 categories)** | **3,728** | **932** | **4,660** |

Table 5: Grasp types and dataset split statistics for TacGlove.

| ID | Grasp Type | Train | Test | Total |
|---|---|---|---|---|
| 0 | Cylindrical | 785 | 207 | 992 |
| 1 | Extension | 64 | 16 | 80 |
| 2 | Parallel 2 | 645 | 151 | 796 |
| 3 | Pinch | 1,006 | 231 | 1,237 |
| 4 | Power Sphere | 246 | 56 | 302 |
| 5 | Quadrupedal | 918 | 255 | 1,173 |
| 6 | Spoon | 31 | 9 | 40 |
| 7 | Writing Tripod | 33 | 7 | 40 |
| | **Total** | **3,728** | **932** | **4,660** |

# B BENCHMARK DETAILS

## B.1 RECONSTRUCTION METRICS

We evaluate reconstruction fidelity using PSNR, MS-SSIM, and Contact F1 on 932 TacGlove test sequences containing 155,109 frames. Reconstructed data are restored to the native taxel ordering before evaluation. Padding positions are excluded from all metrics.

**PSNR.** Reference and reconstructed sequences are normalized using the same sequence-level min-max parameters shared across XYZ. PSNR uses a peak value of one and is computed from the mean squared error aggregated over the test set.

**MS-SSIM.** We compute MS-SSIM (Wang et al., 2003) using three scales with renormalized standard weights to accommodate short sequences. Rate-distortion curves and BD-rate calculations use $-10\log_{10}(1-\text{MS-SSIM})$.

**Contact F1.** A taxel is considered in contact when the absolute value of at least one XYZ component exceeds 0.02 in the original measurement units. We compare reference and reconstructed contact maps (Zhang et al., 2026a) and compute

$$\text{Contact F1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}, \tag{7}$$

where TP counts correctly reconstructed contacts, FP counts spurious contacts, and FN counts missed contacts. These counts are accumulated over all taxels and frames in the test set. The score is set to zero when the denominator is zero.

**Bitrate.** We measure bitrate in Bits Per Taxel (BPT):

$$\text{BPT} = \frac{B}{N\sum_{e=1}^{E} T_e}, \tag{8}$$

where $B$ is the total number of bits in the actual encoded bitstreams, $N = 1{,}207$ is the number of taxels per frame, and $E = 932$ is the number of test sequences. The index $e$ identifies a test sequence, and $T_e$ denotes its number of frames. Padding positions are excluded from the denominator.

## B.2 CLASSIFIER TRAINING AND EVALUATION

For the object and grasp classification tasks introduced in Section 5.1, we train separate LR, RF, and MLP classifiers on the uncompressed training set, treating each complete sequence as one sample. We split the TacGlove dataset into training and evaluation sets at the sequence level in an 8:2 ratio, yielding 3,728 training sequences and 932 evaluation sequences. We divide each sequence into eight temporal intervals and compute the mean and standard deviation of each taxel's XYZ components within each interval to obtain a fixed-length feature vector. The features are standardized, reduced to 128 dimensions using PCA, and standardized again before being passed to the classifier. The preprocessing parameters and classifiers are fitted only on uncompressed training data and remain fixed when evaluating compressed data. Reconstructed sequences undergo the same feature extraction and preprocessing, so changes in accuracy reflect the effect of compression on the classifier inputs.

LR uses balanced class weights, $\ell_2$ regularization with $C = 1$, and L-BFGS with at most 5,000 iterations. RF uses 300 trees and balanced class weights. MLP has one hidden layer with 64 units and dropout 0.1. It is trained with AdamW, a learning rate of $10^{-3}$, a weight decay of $10^{-3}$, a batch size of eight and a 500 epoch budget.

## B.3 BASELINE CONFIGURATIONS

We evaluate ten image codec configurations and three video codec configurations, counting pretrained TCM (PT) and fine-tuned TCM (FT) separately. Table 6 summarizes their operating points. QP denotes the quantization parameter, and $\lambda$ denotes the distortion weight in the rate-distortion training objective.

Table 6: Operating-point settings for the image and video baselines in Section 5.1. TCM (PT) and TCM (FT) use the same six $\lambda$ values. Quality indices and $\lambda$ values are model-specific controls.

| Method | Evaluated settings |
|---|---|
| *Image compression: taxel-time image* | |
| HM / HM-SCC / VTM | QP $\{17, 22, 27, 32, 37, 42\}$ |
| VTM-SCC | QP $\{12, 17, 22, 27, 32, 37, 39, 42\}$ |
| Hyperprior | CompressAI quality indices 1-8 |
| Cheng20 (Cheng2020-anchor) | CompressAI quality indices 1-6 |
| TCM (PT) / TCM (FT) | $\lambda \in \{0.0025, 0.0035, 0.0067, 0.013, 0.025, 0.05\}$ |
| ELIC (reproduction) | Source labels $\lambda \in \{0.004, 0.008, 0.016, 0.032, 0.15, 0.45\}$ |
| LALIC | Released quality indices 1-6 |
| *Video compression: hand-shaped tactile map sequence* | |
| VTM (Inter) | Low-Delay-P; QP $\{17, 22, 27, 32, 37, 42\}$ |
| DCVC-RT | Intra/inter QPs $\{0, 9, 18, 21, 27, 36, 42, 45, 54, 63\}$ |
| DCVC-UF (HT-L) | Intra/inter QPs $\{0, 9, 18, 27, 36, 45, 54, 63\}$ |

**Conventional Codecs.** HM uses HM-16.20 with the main-RExt intra configuration. HM-SCC uses HM-16.20+SCM-8.8 with the main-SCC intra configuration. VTM, VTM-SCC, and VTM (Inter) use VTM-23.3. HM, HM-SCC, VTM, and VTM-SCC use intra coding with 10-bit 4:4:4 inputs, treating XYZ as three component planes. VTM (Inter) uses the Low-Delay-P configuration with equal QPs across the three axes.

**Learned Image Codecs.** Hyperprior and Cheng20 use the MSE-trained checkpoints provided by CompressAI, with Cheng20 implemented as Cheng2020-anchor.[1] TCM (PT)[2] and LALIC[3] use their six author-released checkpoints, while ELIC uses six checkpoints from a public reproduction.[4] All models receive normalized floating-point inputs. Their pretrained weights remain unchanged except for TCM (FT).

**TCM Fine-tuning.** TCM (FT) is initialized from the corresponding pretrained checkpoint at each of the six operating points and fine-tuned on the TacGlove training set for 200 epochs. Each epoch samples 1,024 normalized crops of size $256 \times 256$, with 75% centered on contact regions. The batch size is four.

**Learned Video Codecs.** DCVC-RT and DCVC-UF use the official implementations and pretrained intra/inter checkpoints.[5] DCVC-UF uses the HT-L configuration. Both codecs retain their internal color conversions and receive floating-point inputs through the expected 0-255 interface. Prediction states are reset at sequence boundaries, and each sequence starts with an intra frame. DCVC-RT uses a reset interval of 64 and disables force-zero thresholding (`force_zero_thres=-1`).

[1] CompressAI implementations and pretrained weights: `https://github.com/InterDigitalInc/CompressAI`.

[2] TCM pretrained checkpoints: `https://github.com/jmliu206/LIC_TCM#pretrained-model`.

[3] LALIC implementation and pretrained weights: `https://github.com/sjtu-medialab/RwkvCompress`.

[4] ELIC reproduction and checkpoints: `https://github.com/VincentChandelier/ELiC-ReImplemetation`.

[5] DCVC implementations and pretrained weights: `https://github.com/microsoft/DCVC`.

# C Additional Results

## C.1 Ablation Study of History Frame Length and Network Depth

We conduct ablation studies on the TacGlove test set to examine the history length $L$ and the number of residual blocks $D$, supporting our choice of model configuration.

Table 7: Ablation Study of History Length $L$. **Take $L = 2$ as the anchor of BD-rate (%).**

| $L$ | Params (M) | PSNR | MS-SSIM | Contact F1 |
|---|---|---|---|---|
| 1 | 0.176 | 2.24 | 2.19 | 1.28 |
| 2 | 0.179 | 0.00 | 0.00 | 0.00 |
| 4 | 0.186 | -0.04 | -0.02 | 0.10 |
| 6 | 0.193 | 0.33 | 0.32 | 0.33 |
| 8 | 0.200 | 0.09 | 0.12 | -0.39 |

Table 8: Ablation Study of Network Depth $D$ of Residual Blocks in Mask-guided Entropy Model. **Take $D = 1$ as the anchor of BD-rate (%).**

| $D$ | Params (M) | PSNR | MS-SSIM | Contact F1 |
|---|---|---|---|---|
| 0 | 0.013 | 17.75 | 17.69 | 16.48 |
| 1 | 0.179 | 0.00 | 0.00 | 0.00 |
| 2 | 0.345 | -3.47 | -3.38 | -1.98 |
| 4 | 0.678 | -5.77 | -5.63 | -4.64 |

**History Frame Length.** We fix $D = 1$ and vary $L$ in Table 7. Using only one historical frame increases BD-rate by 2.24% under PSNR, 2.19% under MS-SSIM, and 1.28% under Contact F1. Extending the history from two to four frames yields reductions of only 0.04% and 0.02% under PSNR and MS-SSIM. Longer histories provide no consistent improvement across the three metrics. Although $L = 8$ reduces Contact F1 BD-rate by 0.39%, it slightly increases BD-rate under PSNR and MS-SSIM and requires more parameters. These results support $L = 2$ as a compact choice that captures the main benefits of temporal context.

**Network Depth.** We fix $L = 2$ and vary $D$ in Table 8. Removing all residual blocks increases BD-rate by 17.75% under PSNR, 17.69% under MS-SSIM, and 16.48% under Contact F1. Increasing $D$ to 2 or 4 further improves compression, with PSNR BD-rate reductions of 3.47% and 5.77%. However, the parameter counts rise to 1.93 and 3.78 times that of $D = 1$. We therefore choose one residual block to balance compression performance and model size.

### C.2 Rate Analysis of Mask-Guided Coding

We analyze the bitrate cost and savings of the mask mechanism on the TacGlove test set. Figure 7 compares the bitrate of the mask bitstream with the additional bitrate of the main bitstream when the mask mechanism is removed.

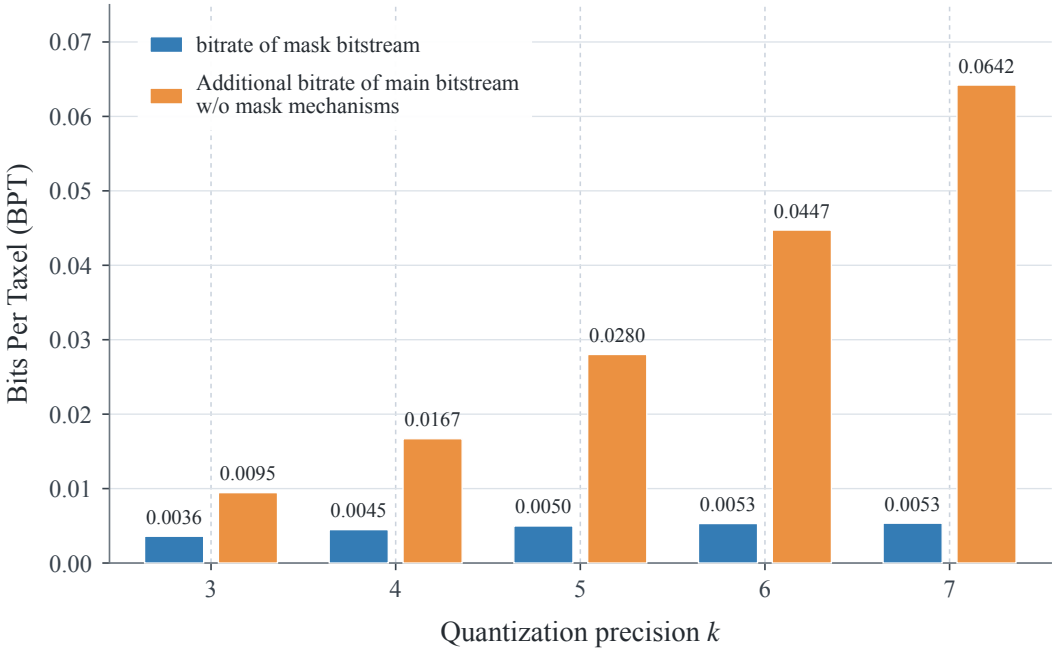


Figure 7: Bitrate of the Mask Bitstream and Additional Bitrate of the Main Bitstream without Mask Mechanisms on the TacGlove Test Set.

The bitrate savings in the main bitstream exceed the mask-stream cost at all evaluated operating points. At $k = 7$, the mask bitstream requires 0.0053 BPT, while the mask mechanism saves 0.0642 BPT in the main bitstream. Thus, transmitting the mask reduces the total bitrate while preserving the same reconstruction quality.