TouchFormer: A Robust Transformer-based Framework for Multimodal Material Perception

Kailin Lyu[1,2], Long Xiao[1,2], Jianing Zeng[1,3], Junhao Dong[4], Xuexin Liu[1], Zhuojun Zou[1], Haoyue Yang[1], Lin Shu[1], Jie Hao[*1]

[1]Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences
[4]Nanyang Technological University

Abstract

Traditional vision-based methods for material perception often experience substantial performance degradation under visually impaired conditions, thereby motivating the shift toward non-visual multimodal material perception. However, current approaches typically fuse modalities naively, overlooking critical challenges including modality-specific noise, the frequent absence of modalities, and their dynamically varying importance across tasks. These limitations lead to suboptimal performance across several benchmark tasks. In this paper, we propose a robust multimodal fusion framework, TouchFormer. Specifically, we employ a Modality-Adaptive Gating (MAG) mechanism and intra- and inter-modality attention mechanisms to adaptively integrate cross-modal features, enhancing model robustness. We further introduce a Cross-Instance Embedding Regularization (CER) strategy to enhance performance in fine-grained subcategory material recognition tasks. Experimental results demonstrate that, compared to existing non-visual methods, the proposed TouchFormer framework achieves classification accuracy improvements of 2.48% and 6.83% on SSMC and USMC tasks, respectively. Additionally, real-world robotic experiments validate TouchFormer's effectiveness in enabling robots to better perceive and interpret their environment, paving the way for its deployment in safety-critical applications such as emergency response and industrial automation.

Website — https://touchformer.github.io/TouchFormer/

1 Introduction

Material perception is a critical capability for both humans and robots when interacting with objects, typically relying on multiple modalities such as vision and touch (Komatsu and Goda 2018; Liu, Sun, and Zhang 2018). Seen Surface Material Classification (SSMC) (Liu et al. 2023; Khojasteh et al. 2024) and Unknown Surface Material Classification (USMC) (Wei et al. 2021) are currently the two most representative material perception tasks, distinguished by whether the task involves recognizing previously unseen categories. Previous work has achieved good performance in scenarios where vision is normal or partially constrained (Tatiya et al. 2024; Khojasteh et al. 2024; Song et al. 2025). However, in scenarios where vision is completely unavailable, such as fire scenes, foggy conditions, or dark factories, the performance of vision-based methods may be affected or even significantly degraded. Therefore, non-visual material perception becomes particularly important.

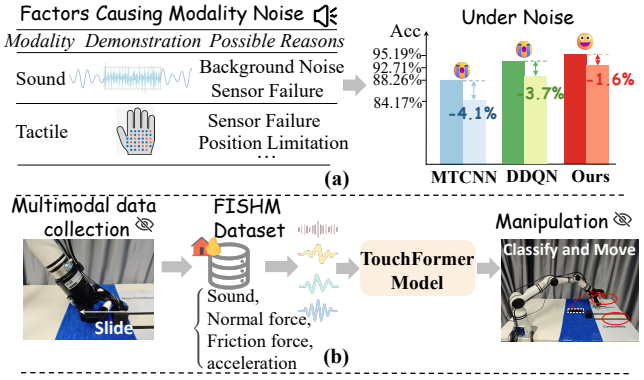

Figure 1: (a) In the SSMC task, when facing real-world modality noise, TouchFormer shows only minor performance loss, in contrast to substantial degradation in baselines like MTCNN (Wei et al. 2021) and DDQN (Liu et al. 2023). (b) Under the simulated fire scenario prototype, the TouchFormer model collects and processes multimodal data (FISHM). Then, leveraging a physics engine to enable robust material perception and manipulation under extreme and vision-constrained conditions.

Although a small number of studies have begun to incorporate multiple non-visual modalities, such as touch and hearing, into material recognition (Liu et al. 2023; Wei et al. 2021), they often overlook the **non-ideality** of data in real-world scenarios. First, sensors across different modalities frequently operate at varied sampling rates, resulting in inherent temporal misalignment between modalities (Tsai et al. 2019). Second, acquiring multimodal data in real-world scenarios inevitably involves noise or sensor failures, which can contaminate or omit information received by the model. This can directly lead to a significant drop in the performance of existing models(Figure 1(a)). Therefore, ensuring the model's robustness to input data in real-world scenarios is essential for handling cases where multimodal in-

*Corresponding author.

puts may be temporally misaligned, partially corrupted, or incomplete (Dong et al. 2022). Moreover, previous multimodal fusion algorithms typically assign equal weights to all modalities and fuse them directly (Wei et al. 2021). However, this is often unreasonable in real-world scenarios. The key modality for identifying different materials often varies (Zhang et al. 2019; Wong, Feng, and Kuo 2023), and assuming equal weights for all modalities may weaken the advantage of the critical modality, thus reducing the general accuracy of the model.

In this paper, we propose a robust multimodal fusion framework, called TouchFormer, designed to address two major issues in existing models: insufficient robustness to input data and suboptimal modality fusion strategies. Specifically, TouchFormer takes noisy or incomplete multimodal sequences as input and employs a Modality-Adaptive Gating (**MAG**) mechanism to dynamically assess the quality of each modality and assign appropriate weights. It then integrates multiple temporally misaligned input modalities through intra-modal and inter-modal attention mechanisms for adaptive fusion based on their importance. As a result, it can effectively extract relevant information from imperfect data to produce enhanced fused modality representations. Building on this, we introduce a Cross-Instance Embedding Regularization (**CER**) strategy to further improve the representational capacity of the embedding features. In conclusion, our main contributions are summarized as follows:

1. To address the issues of insufficient robustness to input data and suboptimal modality fusion in existing non-visual multimodal models, we propose **TouchFormer**, a robust multimodal representation learning framework.
2. We propose three functionally complementary core modules: Modality-Adaptive Gating (MAG), intra- and inter-modal attention mechanisms, and Cross-Instance Embedding Regularization (CER), which together enhance the robustness and representational capacity of TouchFormer.
3. Experimental results demonstrate that TouchFormer outperforms existing non-visual multimodal approaches, achieving accuracy gains of at least **2.48%** on the SSMC task and **6.83%** on the USMC task. In the analysis of robustness to noisy modalities, TouchFormer also achieves the best robustness among existing non-visual multimodal models.
4. To evaluate the model's performance in complex real-world scenarios, we proposed the Fire Incident Sound and Haptic Material **(FISHM)** dataset and simulated a robotic material-sorting task under blindfolded fire conditions. This further validates the effectiveness of the TouchFormer framework in enabling robots to identify materials without relying on vision (Figure 1(b)), demonstrating its promising potential for applications in emergency response and industrial settings.

2 Related Works

2.1 Non-visual Material Perception

In visually impaired environments, touch is the primary perception modality, with robots detecting information such as force and acceleration to sense objects (Calandra et al. 2018; Yuan et al. 2018; Sunil et al. 2023). Additionally, a few studies have utilized audition to distinguish material properties (Shan et al. 2025; Tatiya et al. 2024). Existing work typically integrates multiple modalities (e.g., normal force, friction force, and audition) to leverage their complementarity and enhance perception accuracy (Bhattacharjee et al. 2018; Liu et al. 2023). Wei et al. introduced MTCNN, which integrates energy-spectrum features, dilated convolutions, and sequence pooling into a unified multimodal temporal convolutional network, enabling efficient fusion of acoustic and tactile cues for material recognition (Wei et al. 2021). hojasteh et al. proposed MMUSR, a data-versus-data framework based on the kernel two-sample test that performs material classification on heterogeneous data with minimal manual tuning (Khojasteh et al. 2024). However, existing methods merely perform simple fusion of multiple modalities and require temporal alignment, resulting in poor robustness in real-world scenarios. Our proposed framework allows for input data to be corrupted or partially missing, demonstrating strong potential in real robotic tasks.

2.2 Tactile Sensors

In recent years, tactile sensors have been widely adopted in various robotic applications, including slip detection, object manipulation, insertion, and material recognition (James and Lepora 2020; Dahiya et al. 2013). These sensors can generally be classified into two main categories. The first category comprises vision-based tactile sensors (VBTS) (Dong, Yuan, and Adelson 2017; Lambeta et al. 2020), which capture detailed information about object shape and material properties by observing the deformation of an illuminated membrane. Although these sensors provide high-resolution tactile data, they typically have a limited lifespan and are unsuitable for harsh environments such as fire or disaster scenarios. The second category, represented by uSkin (Funabashi et al. 2018), consists of multi-contact tactile sensors that measure multiple types of signals, including force, vibration, and acceleration, through simple and low-dimensional sensing mechanisms. These sensors are known for their low cost and strong durability (Paulino et al. 2017; Tomo et al. 2016). Considering the demanding requirements for robustness and reliability in field environments, this study employs the uSkin sensor, which is capable of simultaneously measuring both normal and frictional forces.

3 Methodology

Figure 2 illustrates the proposed TouchFormer framework. It comprises three functionally complementary core modules: Modality-Adaptive Gating module **(MAG)**, which enhances input reliability by filtering noisy or irrelevant modalities at the source through dynamic weighting. **Intra- and inter-modal** Transformer fusion module facilitates deep integration across modalities, addressing asynchronous alignment and cross-modal interaction, and cross-instance embedding regularization module **(CER)** further optimizes the prototype space to enhance feature discriminability. Together, the three modules form a complementary system that enhances

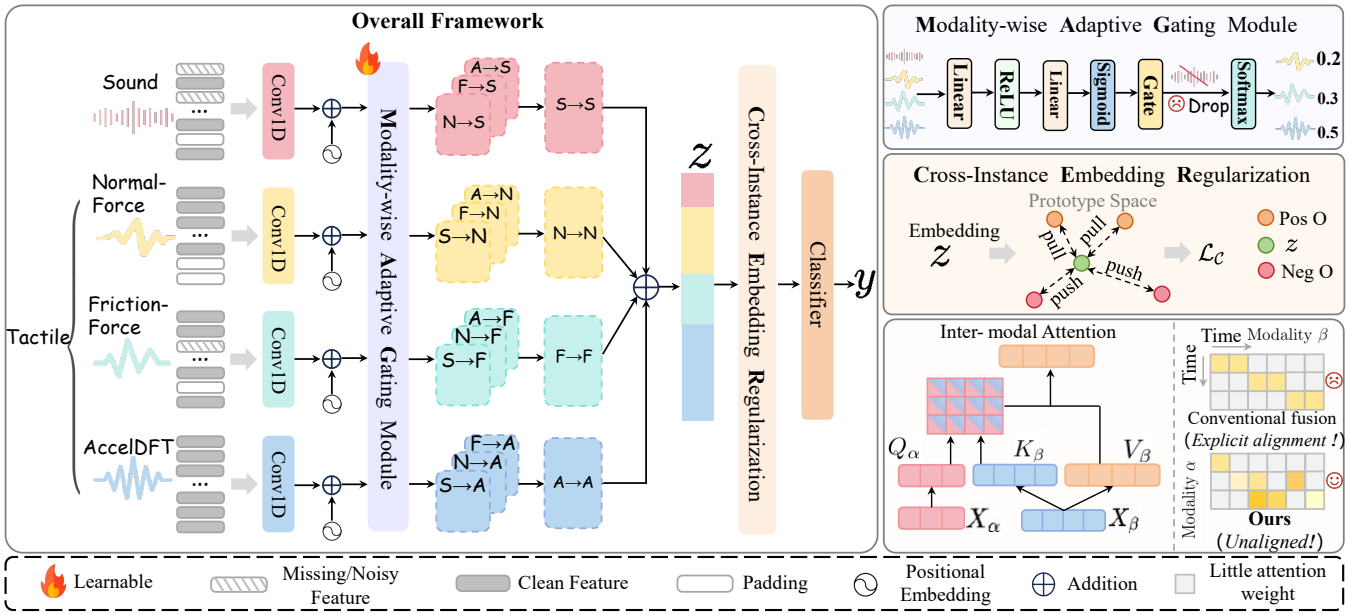

Figure 2: Overview of the proposed framework. TouchFormer receives noisy or incomplete multimodal sequences as input and employs Modality-Adaptive Gating (MAG) to dynamically assess the quality of each modality. It then adaptively integrates cross-modal features within the latent space using both intra- and inter-modality attention mechanisms, without requiring explicit alignment. Finally, Cross-Instance Embedding Regularization (CER) is applied to improve the clarity and discriminability of the representation space, thereby facilitating robust surface material classification.

robustness, **fusion quality**, and **discriminability**, enabling reliable Multimodal material perception in complex real-world environments.

3.1 Modality-Adaptive Gating Module

Multimodal data are inherently noisy and often incomplete, leading to significant disparities in the quality of information provided by different modalities. However, traditional multimodal fusion methods (Khojasteh et al. 2024) typically treat features from different modalities equally, such as by assigning uniform weights or directly concatenating them, without considering modality-specific reliability. This uniform strategy fails to handle biased or noisy inputs, potentially causing negative transfer and performance degradation (Section 1). To address this issue, we propose a **MAG** module, which dynamically evaluates and adaptively adjusts the importance of each modality during feature fusion. Specifically, for each modality $X_m \in \mathbb{R}^{T\times d}$, an intermediate feature representation H_m is first computed via a linear transformation followed by a nonlinear activation function:

$$H_m = \text{ReLU}(W_1 X_m + b_1). \tag{1}$$

Subsequently, modality-specific gating weights $g_m \in [0, 1]$ are generated by applying another linear transformation and a sigmoid activation function to the intermediate representation:

$$g_m = \sigma(W_2 H_m + b_2). \tag{2}$$

We further introduce a hyperparameter $gate_{th}$. Modalities with gating weights below this threshold ($g_m < gate_{th}$) are considered as providing insufficient or noisy information and thus discarded to prevent contamination of the fused multimodal representations.

Finally, to explicitly quantify the relative contributions of each modality(sound S, normal force N, friction force F, and acceleration A) during the fusion process, we apply a softmax normalization on the gating weights to obtain the final modality importance weights α_m:

$$\alpha_m = \frac{\exp(g_m)}{\sum_k \exp(g_k)}, \quad m \in \{S, N, F, A\}. \tag{3}$$

The modality features adjusted through the aforementioned adaptive gating are computed as follows:

$$Z_m = \alpha_m \odot (X_m + PE(T, d)), \tag{4}$$

where $PE(T, d)$ denotes positional embeddings, and $\odot$ denotes the element-wise multiplication operation. This operation enables the model to dynamically adapt to different modalities, filtering out low-quality modalities and improving input reliability, thereby enhancing the robustness of subsequent feature fusion.

3.2 Inter- and Intra-modal Transformer Fusion Module

Conventional multimodal fusion adopts two paradigms: ***i)*** concatenating raw or intermediate features at a fixed layer before a Transformer (Shi et al. 2022; Zheng et al. 2021), and ***ii)*** using convolutional networks to extract modality specific features and then concatenating them (Zhan et al. 2021; Liu et al. 2024). Both strategies generally require manual

alignment of modality sequences to a common time step. In embodied-intelligence perception scenarios, however, heterogeneous **sensors exhibit inherent latency**, making such alignment both **labor-intensive** and **error-prone**. To capture within-modality temporal structures and cross-modality semantic dependencies, we simultaneously model intra- and inter-modal interactions. Inspired by MulT (Tsai et al., 2019), we adopt cross-modal attention to inject low-level source features into target modalities through explicit cross modal attention, enabling robust fusion of asynchronous sequences without alignment. Unlike MulT, our method applies **MAG** again before final feature integration, reweighting block negative transfer, enabling further adaptive fusion and enhancing both flexibility and robustness.

Temporal Convolution and Positional Embedding. To allow each sequence element to perceive its local neighbourhood, we first apply a one–dimensional temporal convolution to the four modalities X_m:

$$\hat{X}_m = \mathrm{Conv1D}\big(X_m, k_m\big) \in \mathbb{R}^{T_m \times d}, \quad m \in \{S, N, F, A\}, \tag{5}$$

where k_m denotes the kernel size of modality m and d is the unified feature dimension. We then add positional embeddings:

$$Z_m^{[0]} = \hat{X}_m + PE(T_m, d). \tag{6}$$

Inter-modal Transformer. Let α be the target modality and β the source modality. We first compute the standard cross-modal attention

$$\hat{Y}_{\alpha \leftarrow \beta} = \mathrm{softmax}\Big(\frac{Q_\alpha K_\beta^\top}{\sqrt{d_k}}\Big) V_\beta, \tag{7}$$

The attention output is subsequently modulated by the modality importance weight α_β, which is computed by the **MAG** module (Section 3.1), such that $Y_{\alpha \leftarrow \beta} = \alpha_\beta \hat{Y}_{\alpha \leftarrow \beta}$. $\alpha_\beta \in [0, 1]$ reflects the reliability of the source modality β; a higher weight indicates a greater contribution to the representation of the target modality.

Intra-modal Transformer. The inter-modal output is merged with the original representation by a residual connection:

$$\tilde{Z}_\alpha = Z_\alpha^{[0]} + Y_{\alpha \leftarrow \beta}. \tag{8}$$

Self-attention is then applied within the same modality:

$$Z_\alpha^{\mathrm{intra}} = \mathrm{Transformer}\big(\tilde{Z}_\alpha, \tilde{Z}_\alpha, \tilde{Z}_\alpha\big). \tag{9}$$

Finally, the intra-modal representations of the four modalities, Z_m^{intra}, are re-weighted by their importance coefficients α_m and concatenated to form the fused feature vector:

$$Z = \mathrm{Concat}\big[\alpha_S Z_S^{\mathrm{intra}}, \alpha_N Z_N^{\mathrm{intra}}, \alpha_F Z_F^{\mathrm{intra}}, \alpha_A Z_A^{\mathrm{intra}}\big]. \tag{10}$$

This sequence jointly captures intra-modal temporal dynamics and inter-modal interactions, enabling the model to learn rich, complementary features for robust multimodal representation.

3.3 Cross-Instance Embedding Regularization

In multimodal representation learning, despite improved information integration through modality selection and fusion strategies, models may still exhibit inter-class confusion and intra-class scattering (Zhou et al. 2023; Jiang et al. 2023). To address these limitations and enhance both the discriminative capacity and generalization of learned representations, we introduce a Cross-Instance Embedding Regularization **(CER)** module. Grounded in contrastive learning principles, CER enforces structural constraints on the embedding space from a cross-instance perspective by leveraging global supervision signals. It promotes intra-class compactness and inter-class separability, thereby enhancing the clarity and discriminability of the overall representation space.

Given a batch of N samples with ℓ_2-normalized embeddings $\{z_i\}_{i=1}^N$ and corresponding labels $\{y_i\}_{i=1}^N$, we construct the similarity matrix $S_{ij} = z_i^\top z_j$. The cross-instance contrastive loss $\mathcal{L}_C$ is defined as

$$\mathcal{L}_C = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\sum_{j \neq i} \mathbf{I}_{\{y_i = y_j\}} \exp\big(S_{ij}/\tau\big)}{\sum_{j \neq i} \exp\big(S_{ij}/\tau\big)}, \tag{11}$$

where τ is the temperature hyper-parameter and $\mathbf{I}_{\{y_i = y_j\}}$ is an indicator function that equals 1 only for positive pairs $(y_i = y_j)$. This loss pulls together embeddings from the same class while pushing apart those from different classes, encouraging a discriminative feature space. As it operates on relative distances rather than fixed labels, it is well suited for tasks requiring instance-level semantic mapping (Section 5.2).

3.4 Overall Loss

The model is trained in a single stage by jointly minimizing the classification loss $\mathcal{L}_{cls}$ and the CER loss $\mathcal{L}_C$:

$$\mathcal{L}_{total} = \mathcal{L}_{cls}(y, \hat{y}) + \lambda \mathcal{L}_C, \tag{12}$$

where $\hat{y}$ denotes the predicted labels, y the ground-truth labels, and λ is a weighting factor that controls the influence of $\mathcal{L}_{total}$ on the total objective. The parameters of the **MAG** module are updated concurrently with the network. Minimizing $\mathcal{L}_C$ enhances both the robustness of the multimodal representations and the model's ability to generalize in classification tasks.

4 Experiment Setup

4.1 Datasets

LMT Haptic Material Database (LMTHM): This publicly available multimodal dataset was collected by Strese et al. (Strese et al. 2019) using the self-developed Texplorer2 device. The dataset comprises 965 samples from 193 distinct surface materials, with five samples per material. It includes multimodal data such as sound, acceleration, normal force, and frictional force. The materials are categorized into eight major classes and several subclasses.

Fire Incident Sound and Haptic Material (FISHM) Dataset Figure 3, acquired through a custom multimodal tactile fingertip. The integrated fingertip employs a uSkin tactile sensor for Normal and Frictional force measurement, a microphone for Sound measurement, and a 12-DoF IMU

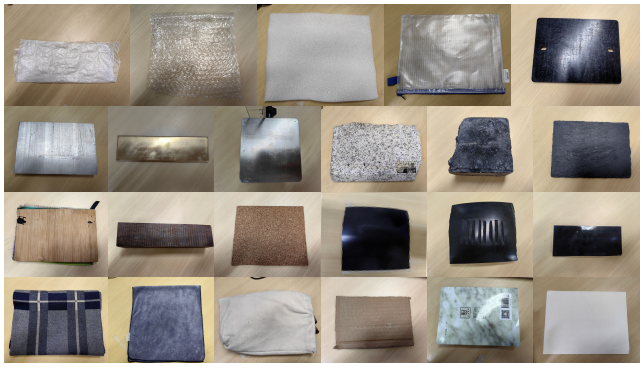

Figure 3: The FISHM dataset comprises seven distinct categories of daily objects, encompassing representative item types potentially encountered in fire incidents.

for tri-axial Acceleration measurement, enabling simultaneous collection of four modalities. It comprises 7 daily material categories, with metal and stone classified as *Non-combustible* and others as *Combustible*. Furthermore, it accurately simulates fire scenarios, thereby facilitating a more effective evaluation of the robustness and generalizability of the proposed model.

4.2 Baselines

We carefully selected widely used state-of-the-art robotic sensing techniques based on the uSkin tactile sensor for multimodal surface material classification tasks to compare with our proposed method, TouchFormer. Specifically, for methods **excluding the visual modality**, we selected LSTM (Ji et al. 2015), MTCNN (Wei et al. 2021), and DDQN (Liu et al. 2023) as baseline models, each capable of classifying surface materials solely using auditory and multiple tactile modalities (including acceleration, normal force, and frictional force). For methods **incorporating the visual modality**, we chose the recently proposed MMUSR (Khojasteh et al. 2024), which uses up to nine different sensor modalities (e.g. visual and tactile signals) for classification.

4.3 Implementation Details

All experiments were conducted on machines equipped with A100 GPUs using the PyTorch framework. During training, we used a batch size of 32 and the Adam optimizer with a weight decay of 0.1. The initial learning rate was set to 0.1 and gradually decayed to 0 using a cosine annealing strategy (Loshchilov and Hutter 2016). The models were trained for a total of 50 epochs. In the robotic application phase, the model trained on the LMTHM dataset was deployed on a robotic arm and fine-tuned using the FISHM dataset for domain adaptation. For the USMC task, we adopted a five-fold cross-validation strategy, consistent with the evaluation protocol used by MTCNN (Wei et al. 2021), to ensure a fair comparison across methods. For the SSMC task, we split the dataset into training and testing sets at a 7:3 ratio. In both settings, performance was evaluated using the mean classification accuracy and G_mean (Espíndola and Ebecken 2005) as the primary metrics.

5 Experimental Results

In this section, we conduct experiments based on the dataset collected using the uSkin sensor to evaluate the performance of our model in surface material classification under various conditions. The objectives are as follows: ***i).*** To verify whether the proposed TouchFormer framework outperforms baseline methods and to demonstrate the effectiveness of each component within the TouchFormer framework. ***ii).*** To assess the model's robustness under randomly corrupted modality features. ***iii).*** To evaluate the framework's performance in enabling robotic environmental understanding in real-world physical scenarios.

Task	Class num	Method	Accuracy [%]	G_mean
SSMC	8	DDQN	92.71	0.93
		MMUSR	79.7	0.80
		MMUSR*	93.8	0.94
		Ours	**95.19** (↑ 2.48)	**0.95**
USMC	8	LSTM†	74.23	0.73
		DALNet†	46.67	0.47
		MTCNN†	87.55	0.88
		Ours	**94.38** (↑ 6.83)	**0.94**
SSMC (Fine-Grained)	193	LSTM	65.75	0.66
		DALNet	42.62	0.43
		MTCNN	80.21	0.80
		Ours	**89.77** (↑ 9.56)	**0.90**

Table 1: Multimodal surface material classification performance on the LMTHM dataset. "$\star$" indicates that the model employs visual input. "†" denote the corresponding sources of the reported results (Wei et al. 2021).

Multimodal Inputs				LSTM	MTCNN	**Ours**
S	***N***	***F***	***A***			
✗	✓	✓	✓	73.78%	83.43%	**87.92%**
✓	✗	✓	✓	71.69%	83.18%	**88.55%**
✓	✓	✗	✓	69.09%	84.99%	**89.12%**
✓	✓	✓	✗	75.50%	87.00%	**89.84%**
✓	✓	✓	✓	74.23%	87.55%	**94.38%**

Table 2: Comparison of classification accuracy as sound (S), normal force (N), friction force (F), and acceleration (A) modalities are incrementally removed. Even with any single modality missing, TouchFormer consistently outperforms other methods using full-modal inputs.

5.1 Multimodal Performance Comparison

We investigated whether our proposed method outperforms existing benchmark approaches under multimodal conditions and assessed the necessity and effectiveness of integrating auditory and multi-tactile data for the surface material classification task. The results presented in Table 1 demonstrate that, in multimodal scenarios, the TouchFormer

framework achieves classification performance improvements of 2.48% and 6.83% over the best baseline methods on the SSMC and USMC tasks, respectively. Notably, even with no visual modality input, our method still surpasses the MMUSR (Khojasteh et al. 2024) includes visual modality input and achieves an accuracy close to the 97.2% performance obtained by MMUSR, which integrates a total of eight modalities derived from six different sensors, including a camera, multiple tactile sensors, an infrared(IR) surface reflectance sensor and a metal detection sensor. Additionally, Table 2 shows that in the USMC task, even after individually removing modalities such as sound, normal force, friction force, and acceleration, TouchFormer still outperforms other methods.

5.2 Fine-Grained Subclass Classification

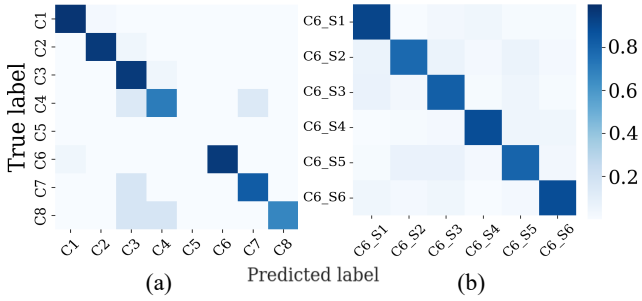

Figure 4: TouchFormer confusion matrix performance on the SSMC task. (a) Coarse-grained classification results across 8 material classes. (b) Fine-grained classification results (e.g., using category C6 as an example).

Human tactile perception relies on both coarse-grained sensing and fine-grained discrimination primarily enabled by the fast-adapting (FA) and slow-adapting (SA) systems (Park et al. 2016). In addition to evaluating our method on a coarse-level surface material classification task with eight categories, we introduce a novel, more challenging fine-grained subclass classification task, such as distinguishing *softwood* from *hardwood* within the broader category of *wood*, which previous research has not addressed. As shown in Table 1 and Figure 4, our proposed method achieves outstanding classification performance across all categories in this task. It is primarily attributed to the proposed MAG and CER, which improve classification accuracy by adaptively selecting and integrating the most relevant modalities and enhancing subclass discriminability within the prototype embedding space (Figure 5).

5.3 Model Robustness for Modality Anomaly

We investigate the robustness of the TouchFormer framework under varying levels of noise applied to randomly selected modalities. Specifically, during both training and testing phases, we introduce noise to randomly selected modalities under different gaussian noise intensities, with a corruption ratio $p \in \{0.0, 0.1, \ldots, 1.0\}$. As shown in Figure 6, the TouchFormer framework consistently outperforms baseline methods across various noise combinations and intensity

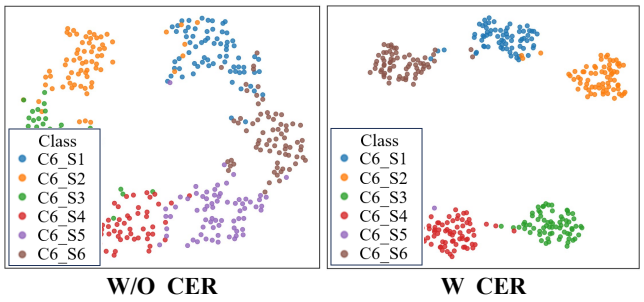

Figure 5: Visualization of subclass embeddings with and without CER. t-SNE plots of feature embeddings for coarse- and fine-grained classification.

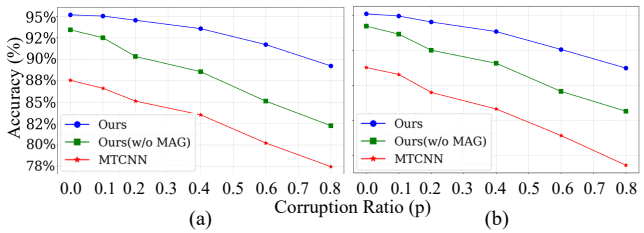

Figure 6: Performance under modality corruption. We compare the classification performance of TouchFormer and other method under varying corruption ratios p.

Configurations	SSMC	USMC	Fine-Grained
Baseline	91.32%	90.17%	83.53%
Baseline + MAG	93.15%	92.56%	84.88%
Baseline + MAG + CER	**95.19%**	**94.38%**	**89.77%**

Table 3: Ablation study of different components across diverse tasks

levels, demonstrating strong robustness in multimodal noisy environments. This improvement is primarily attributed to the **MAG** module, which dynamically evaluates and adjusts the contribution of each modality during feature fusion.

5.4 Ablation Study

Table 3 presents the ablation study results of the proposed TouchFormer model on the SSMC, USMC, and fine-grained classification tasks. The baseline refers to a configuration that employs only intra- and inter-modal Transformer fusion modules without any of the proposed enhancements. We first incorporate the MAG module into the baseline, which aims to adaptively select and fuse the most relevant modalities. This step improves performance from (91.32, 90.17, 83.53) to (93.15, 92.56, 84.88). Next, we introduce CER to enhance the discriminability of different categories in the prototype embedding space. After applying CER, classification accuracy improves by (+2.04%, +1.82%, +4.89%). The corresponding visualization results are shown in Figures 5 and 4. We conduct further experiments by adjusting $gate_{th}$ and λ to evaluate their impact on TouchFormer's performance.

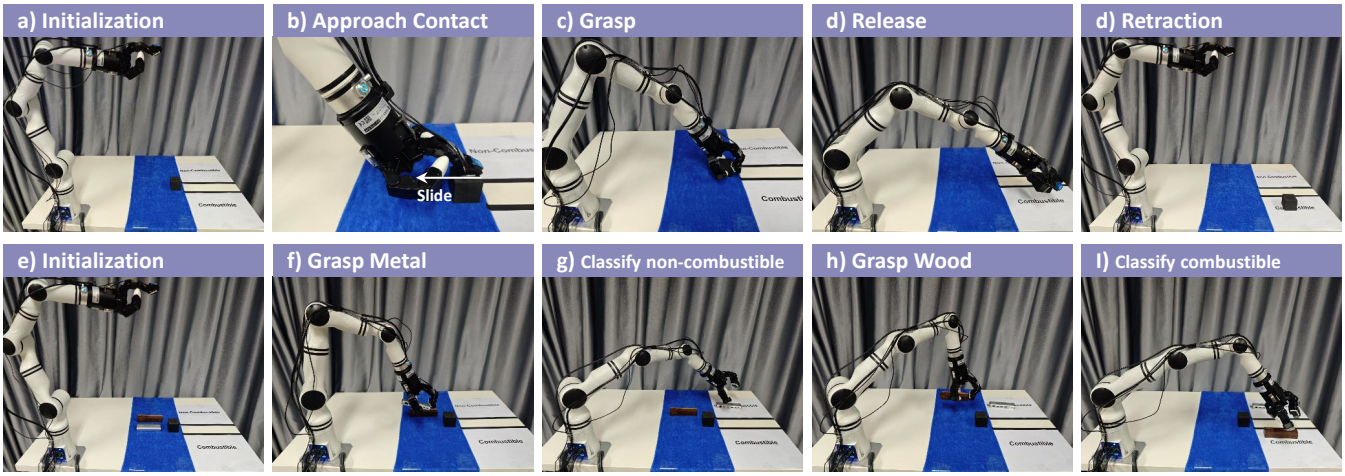

Figure 7: Evaluation of material transport performance using the RealMan robot in a vision-free, noise-disturbed simulated fire scenario. (a)∼(d) illustrate the procedure for evaluating individual objects; (e)∼(i) depict the evaluation process in multi-object scenes. The robot identifies the material properties of wood, rubber, and metal blocks, and executing appropriate grasp-and-place actions to move them into their corresponding zones.

5.5 Robotic Application

Robotic material perception, such as surface material classification, aims to enhance robots' ability to understand and interact with their environments. While prior experiments validated our method's effectiveness, a gap remains between perception and manipulation. To address this, we present a fire-scenario prototype demonstrating how material perception models guide strategy inference for robotic interaction.

We apply material perception techniques to a robotic sorting task in a simulated fire scenario, using a Realman RM65-B 6-DoF robotic arm equipped with a TESOLLO Gripper-3F three-fingered dexterous hand to perform object grasping and releasing operations. This task simulates key complexities inherent in fire environments, characterized by three principal constraints: compromised visibility, confined spatial configurations, and multiple objects. The procedure is as follows: (1) the robot system operates in a multi-object, low-visibility environment where combustible and non-combustible objects remain fixed at predetermined positions, with designated collection zones on the lateral sides; (2) using multimodal perception, the robot identifies and classifies materials to determine their flammability; and (3) Based on the classification results, the robot executes a predefined program to move each object to its designated area while clearing the safety pathway. Since no visual input is available, object manipulation and grasping during the experiment are executed using fixed-parameter programs. The object category is determined solely by the algorithm. The setup is illustrated in Figure 7. We evaluate the performance of TouchFormer on various tasks using the FISHM dataset. The experimental results are summarized in Table 4. In addition, we assess the model's real-world performance by integrating it with a physics engine. Specifically, we measure the classification accuracy of the robot system in identifying combustible and non-combustible materials within the designated collection zones during the SSMC task.

Environments	SSMC	USMC	Fine-Grained
w/o Noise	91.47%	90.26%	90.05%
w/ Noise	89.54%	88.03%	87.76%

Table 4: Performance of the TouchFormer on the FISHM dataset across different environments and tasks.

6 Conclusion

TouchFormer introduces a novel and robust framework for multimodal material perception under visually impaired and noisy conditions, achieving superior performance via adaptive and noise-aware fusion. The framework integrates a Modality-Adaptive Gating (MAG) mechanism alongside intra- and inter-modal Transformer-based fusion to explicitly account for modality noise and missing inputs, which are prevalent in real-world settings. Furthermore, the proposed Cross-Instance Embedding Regularization (CER) strategy enhances feature discriminability, particularly for fine-grained material classification. Extensive experiments demonstrate that TouchFormer consistently outperforms existing vision-free baselines across multiple benchmarks. Additionally, real-world robotic experiments confirm its capability to support accurate environmental understanding and informed action inference, underscoring its strong potential for deployment in high-stakes scenarios such as emergency response and industrial automation.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62236007.

References

Bhattacharjee, T.; Clever, H. M.; Wade, J.; and Kemp, C. C. 2018. Multimodal tactile perception of objects in a real home. *IEEE Robotics and Automation Letters*, 3(3): 2523–2530.

Calandra, R.; Owens, A.; Jayaraman, D.; Lin, J.; Yuan, W.; Malik, J.; Adelson, E. H.; and Levine, S. 2018. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3(4): 3300–3307.

Dahiya, R. S.; Mittendorfer, P.; Valle, M.; Cheng, G.; and Lumelsky, V. J. 2013. Directions toward effective utilization of tactile skin: A review. *IEEE Sensors Journal*, 13(11): 4121–4138.

Dong, J.; Wang, Y.; Lai, J.-H.; and Xie, X. 2022. Improving adversarially robust few-shot image classification with generalizable representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9025–9034.

Dong, S.; Yuan, W.; and Adelson, E. H. 2017. Improved gelsight tactile sensor for measuring geometry and slip. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 137–144. IEEE.

Espíndola, R. P.; and Ebecken, N. F. 2005. On extending f-measure and g-mean metrics to multi-class problems. *WIT Transactions on Information and Communication Technologies*, 35: 25–34.

Funabashi, S.; Morikuni, S.; Geier, A.; Schmitz, A.; Ogasa, S.; Torno, T. P.; Somlor, S.; and Sugano, S. 2018. Object recognition through active sensing using a multi-fingered robot hand with 3d tactile sensors. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2589–2595. IEEE.

James, J. W.; and Lepora, N. F. 2020. Slip detection for grasp stabilization with a multifingered tactile robot hand. *IEEE Transactions on Robotics*, 37(2): 506–519.

Ji, M.; Fang, L.; Zheng, H.; Strese, M.; and Steinbach, E. 2015. Preprocessing-free surface material classification using convolutional neural networks pretrained by sparse autoencoder. In *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6. IEEE.

Jiang, Q.; Chen, C.; Zhao, H.; Chen, L.; Ping, Q.; Tran, S. D.; Xu, Y.; Zeng, B.; and Chilimbi, T. 2023. Understanding and constructing latent modality structures in multi-modal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7661–7671.

Khojasteh, B.; Solowjow, F.; Trimpe, S.; and Kuchenbecker, K. J. 2024. Multimodal Multi-User Surface Recognition With the Kernel Two-Sample Test. *IEEE Trans Autom. Sci. Eng.*

Komatsu, H.; and Goda, N. 2018. Neural mechanisms of material perception: Quest on Shitsukan. *Neuroscience*, 392: 329–347.

Lambeta, M.; Chou, P.-W.; Tian, S.; Yang, B.; Maloon, B.; Most, V. R.; Stroud, D.; Santos, R.; Byagowi, A.; Kammerer, G.; et al. 2020. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3): 3838–3845.

Liu, G.; Lv, S.; Wang, C.; Li, X.; and Nai, W. 2023. Surface material classification based on unbalanced visual and haptic data: A double-DQN method. *IEEE Transactions on Instrumentation and Measurement*, 72: 1–12.

Liu, H.; Sun, F.; and Zhang, X. 2018. Robotic material perception using active multimodal fusion. *IEEE Transactions on Industrial Electronics*, 66(12): 9878–9886.

Liu, Z.; Chi, C.; Cousineau, E.; Kuppuswamy, N.; Burchfiel, B.; and Song, S. 2024. Maniwav: Learning robot manipulation from in-the-wild audio-visual data. In *8th Annual Conference on Robot Learning*.

Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Park, S. B.; Davare, M.; Falla, M.; Kennedy, W. R.; Selim, M. M.; Wendelschafer-Crabb, G.; and Koltzenburg, M. 2016. Fast-adapting mechanoreceptors are important for force control in precision grip but not for sensorimotor memory. *Journal of neurophysiology*, 115(6): 3156–3161.

Paulino, T.; Ribeiro, P.; Neto, M.; Cardoso, S.; Schmitz, A.; Santos-Victor, J.; Bernardino, A.; and Jamone, L. 2017. Low-cost 3-axis soft tactile sensors for the human-friendly robot Vizzy. In *2017 IEEE international conference on robotics and automation (ICRA)*, 966–971. IEEE.

Shan, Q.; Cao, Y.; Chi, H.; Fan, S.; Zhu, Z.; and Hou, D. 2025. A Star-Nose-Inspired Bionic Soft Robot for Nonvisual Spatial Detection and Reconstruction. *Advanced Intelligent Systems*, 7(1): 2400601.

Shi, B.; Hsu, W.-N.; Lakhotia, K.; and Mohamed, A. 2022. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*.

Song, W.; Zhou, Z.; Zhao, H.; Chen, J.; Ding, P.; Yan, H.; Huang, Y.; Tang, F.; Wang, D.; and Li, H. 2025. Reconvla: Reconstructive vision-language-action model as effective robot perceiver. *arXiv preprint arXiv:2508.10333*.

Strese, M.; Brudermueller, L.; Kirsch, J.; and Steinbach, E. 2019. Haptic material analysis and classification inspired by human exploratory procedures. *IEEE transactions on haptics*, 13(2): 404–424.

Sunil, N.; Wang, S.; She, Y.; Adelson, E.; and Garcia, A. R. 2023. Visuotactile affordances for cloth manipulation with local control. In *Conference on Robot Learning*, 1596–1606. PMLR.

Tatiya, G.; Francis, J.; Wu, H.-H.; Bisk, Y.; and Sinapov, J. 2024. Mosaic: Learning unified multi-sensory object property representations for robot learning via interactive perception. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 15381–15387. IEEE.

Tomo, T. P.; Wong, W. K.; Schmitz, A.; Kristanto, H.; Sarazin, A.; Jamone, L.; Somlor, S.; and Sugano, S. 2016. A modular, distributed, soft, 3-axis sensor system for robot hands. In *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, 454–460. IEEE.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, 6558.

Wei, J.; Cui, S.; Hu, J.; Hao, P.; Wang, S.; and Lou, Z. 2021. Multimodal unknown surface material classification and its application to physical reasoning. *IEEE Transactions on Industrial Informatics*, 18(7): 4406–4416.

Wong, C.-C.; Feng, H.-M.; and Kuo, K.-L. 2023. Multisensor fusion simultaneous localization mapping based on deep reinforcement learning and multi-model adaptive estimation. *Sensors*, 24(1): 48.

Yuan, W.; Mo, Y.; Wang, S.; and Adelson, E. H. 2018. Active clothing material perception using tactile sensing and deep learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 4842–4849. IEEE.

Zhan, X.; Wu, Y.; Dong, X.; Wei, Y.; Lu, M.; Zhang, Y.; Xu, H.; and Liang, X. 2021. Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11782–11791.

Zhang, J.; Wu, Y.; Feng, W.; and Wang, J. 2019. Spatially attentive visual tracking using multi-model adaptive response fusion. *Ieee Access*, 7: 83873–83887.

Zheng, R.; Chen, J.; Ma, M.; and Huang, L. 2021. Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation. In *International Conference on Machine Learning*, 12736–12746. PMLR.

Zhou, X.; Zhou, D.; Hu, D.; Zhou, H.; and Ouyang, W. 2023. Exploiting visual context semantics for sound source localization. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 5199–5208.