Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Fusion-VTT: Visual-Tactile-Text Fusion Learning for Robotic Object Recognition

Li Yang
Huazhong University of Science and Technology
Liangqi Zhang
Wuhan Huaweike Intelligent Technology Co., Ltd.
Min Li
min.li@hust.edu.cn

Huazhong University of Science and Technology

Research Article

Keywords: Multimodal integration, object recognition, tactile perception, visual perception

Posted Date: March 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9067694/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Fusion-VTT: Visual-Tactile-Text Fusion Learning for Robotic Object Recognition

Li Yang[1], Liangqi Zhang[2], Min Li[1,2*]

[1*]State Key Laboratory of Intelligent Manufacturing Equipment and Technology, Huazhong University of Science and Technology, Wuhan, China.

[2]Wuhan Huaweike Intelligent Technology Co., Ltd., Wuhan, China.

*Corresponding author(s). E-mail(s): min.li@hust.edu.cn;

Abstract

Multimodal fusion is a promising approach to enhance environmental perception and object recognition for robotic systems. However, the inherent heterogeneity and semantic discrepancies among visual, tactile, and textual modalities pose significant challenges for feature fusion. This paper proposes a novel hierarchical fusion framework, Fusion-VTT, designed to achieve deep feature-level fusion across visual, tactile, and textual modalities. The framework first employs Patch Embedding and a Pre-trained Bidirectional Encoder Representations from Transformers (BERT) model to extract features from three modalities, followed by spatial alignment. It then innovatively incorporates parameter-sharing cross-attention mechanism to establish six bidirectional interaction pathways between modality pairs, thereby capturing fine-grained cross-modal correlations. Subsequently, a self-attention module is applied to integrate features into a globally consistent representation. To evaluate the proposed method, a new multimodal dataset comprising 20 categories of common household objects was compiled. Experimental results on this custom dataset and the public MSDO dataset demonstrate that Fusion-VTT achieves a recognition accuracy of 99.23%, substantially outperforming existing baseline methods and confirming the effectiveness of the proposed fusion strategy.

Keywords: Multimodal integration, object recognition, tactile perception, visual perception

1 Introduction

Significant progress has been made in the ability of robotic systems to interact with humans and complex environments over the past decade. When interacting with the surrounding environment and performing complex and delicate tasks, precise and efficient environmental perception and object recognition are very critical for robotic systems. Vision-based methods are one of the most commonly used approaches for object recognition[1] due to varying advantages, such as non-contact and high efficiency[2]. However, the performance of vision-based perception systems is limited in non-ideal scenarios such as lighting changes and partial occlusion[3]. Meanwhile, tactile sensing technology, utilizing sensors such as force-sensitive resistor arrays[4] and GelSight tactile imaging devices[5], enables the precise capture of mechanical properties, including material hardness and surface texture, thereby complementing the limitations of visual perception[6]. However, tactile sensing inherently depends on physical contact, which constrains its data acquisition efficiency[7].

From the perspective of the perceptual mechanism, humans do not rely on a single sensory channel to understand their environment or infer the physical properties of objects. Instead, people integrate visual cues such as shape and color, tactile experiences including texture and hardness, and semantic descriptions to form a comprehensive and accurate environmental perception[8]. This integrative capability offers valuable inspiration for robotic perception systems. By fusing information from multiple modalities—such as vision, touch, text, and sound—robots could achieve a more holistic environmental awareness and make more intelligent decisions[9]. The incorporation of tactile feedback is especially critical in applications requiring fine manipulation or occurring in visually degraded conditions, such as robotic surgery and object handling in cluttered settings[10]. Furthermore, the ability to interpret and utilize semantic text allows robots to interact more naturally with humans and perform tasks requiring language understanding[11]. Combining visual and textual information could help bridge the gap between pixel-level data and high-level semantics, leading to more accurate environmental interpretation and task execution[12].

Nevertheless, multimodal fusion still faces two core challenges. First, data from visual, tactile, and textual modalities are inherently heterogeneous, complicating direct feature integration[13]. Visual data typically consists of spatially structured RGB images, tactile data often takes the form of time-series pressure sequences, and text comprises discrete symbolic representations. These differences result in misaligned statistical distributions and semantic gaps, posing substantial difficulties for fusion algorithms. Second, high-quality multimodal datasets remain scarce, and many existing collections lack precise cross-modal alignment, which hinders effective training of fusion models[14]. Although several datasets have been introduced (e.g., GelFabric[15], LMT[16], Vitac[17]), they focus on specific material types such as fabrics, leaving a gap in datasets encompassing commonly seen objects. This limitation curtails the broader application of multimodal perception technologies in robotics.

Inspired by the multimodal perceptual mechanism observed in humans, this paper proposes a novel tri-modal fusion framework that integrates visual, tactile, and textual information to significantly enhance object recognition accuracy through deep feature-level integration. The proposed approach adopts a hierarchical fusion architecture. To

facilitate evaluation, a new multimodal dataset was constructed, comprising 20 categories of common daily objects. Comprehensive experiments using both self-collected datasets and public benchmarks demonstrate that the proposed fusion framework achieves superior classification performance compared to previous methods, confirming its effectiveness in multimodal information processing. The remainder of this paper is organized as follows: Section 2 reviews previous and related work; Section 3 details the model architecture; Section 4 explains the data collection and processing procedures, followed by Section 5 presenting experimental results and comparisons; in the end, Section 6 concludes the study and suggests future research directions.

2 Related Work

In robotic perception, visual systems play a crucial role. By processing visual information, robots could quickly and accurately detect environmental changes[18]. However, visual data primarily captures surface attributes and offers limited insight into intrinsic physical properties such as material composition and hardness[19]. In recent years, tactile sensing technology has advanced significantly as a key means for robots to interact with their environment. Various tactile sensors have been integrated into robotic end-effectors and biomimetic prostheses[20], enabling critical functions including object slippage detection[21], tactile object recognition[22], and material surface classification[23].

Nevertheless, in complex unstructured environments, unimodal perception shows significant limitations. Visual perception, for example, is often compromised under low-light conditions or in the presence of occlusions, whereas tactile sensing depends on physical contact and cannot provide object information beyond direct reach. Thus, multimodal fusion has emerged as an essential direction for advancing robotic environmental understanding. In scenarios where visual information alone is insufficient for accurate object identification or pose estimation, supplementary sensory inputs could substantially enhance robots' ability to interpret and interact with objects[13]. By integrating complementary information from heterogeneous sources such as vision, touch, and language, multimodal fusion methods markedly improve the robustness and accuracy of perceptual models[24].

In visual-tactile fusion research, Takahashi et al.[25] adopted an encoder-decoder architecture to map visual and tactile features into a shared latent space, facilitating the recognition of unknown materials. Yuan et al.[15] employed GelSight sensors to construct a unified visual-tactile representation, enhancing robotic perception in fabric material recognition. Zhang et al.[26] proposed a Partial Visual-Tactile Fusion (PVTF) framework that reduces inter-modal discrepancies and exploits complementary features, improving both object recognition accuracy and model robustness under incomplete inputs. Despite these advances, such methods remain confined to physical modalities and do not incorporate higher-level semantic information, thus limiting their applicability in tasks requiring semantic interaction.

Advances in Natural Language Processing (NLP) and Computer Vision (CV) have made textual modality integration an important avenue for enhancing perceptual and reasoning capabilities in multimodal systems. The Contrastive Language - Image

Pretraining (CLIP) model, introduced by Radford et al.[27], aligns images and text through contrastive learning and demonstrates strong performance in zero-shot recognition tasks. Huang et al.[28] designed an end-to-end multimodal fusion network that integrates scene understanding, leading to improved system performance. Although these methods have made substantial progress in visual-text fusion, they have not effectively incorporated physical modalities such as tactile sensing, limiting their use in scenarios requiring physical interaction, such as robotic manipulation.

Multimodal fusion research is increasingly expanding from bimodal to more complex multimodal settings. Lin et al.[29] drew inspiration from CLIP to map text, audio, and images into a common representation space, achieving sample-efficient cross-modal learning. Girdhar et al.[30] proposed ImageBind, which integrates six modalities for joint embedding learning and validated its efficacy through cross-modal retrieval and zero-shot tasks. While those general-purpose models illustrate the potential of multimodal fusion, their broad scope often results in high computational costs and inadequately compact feature representations for specific applications.

To address the aforementioned challenges, this study focuses on the modality heterogeneity in visual-tactile-text fusion. A feature space alignment method is proposed to bridge the representational gaps between modalities, along with a cross-attention-based fusion architecture to enhance multimodal feature interaction. Meanwhile, a new multimodal dataset has been constructed, comprising 20 object categories with comprehensive visual, tactile, and textual descriptions, thereby providing a robust foundation for model training and evaluation.

3 Method

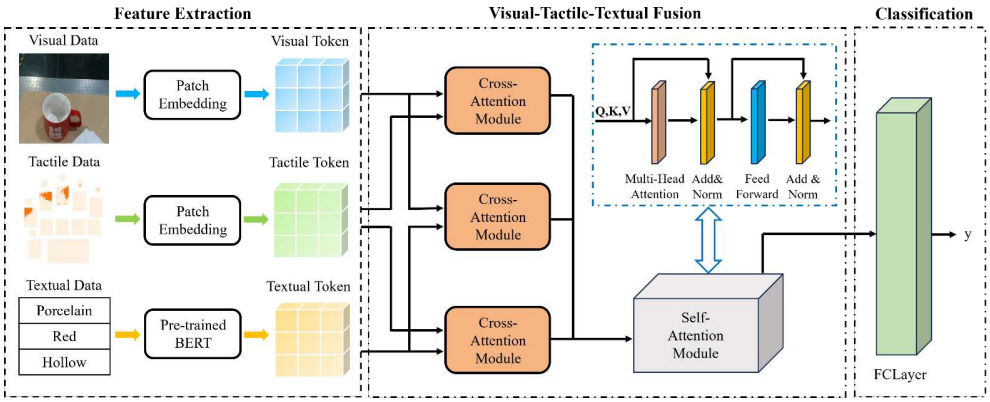


Fig. 1 Schematics of the Model Architecture. The framework comprises three stages: feature extraction, multimodal fusion, and classification. Left: Visual and tactile images are tokenized via Patch Embedding, while text is encoded by a pre-trained BERT model. Middle: Cross-Attention mechanism aligns features across visual, tactile, and textual modalities, followed by a Self-Attention module for global integration. Right: The fused representation is processed by a Fully Connected Layer to predict the object category.

An innovative multimodal feature fusion model for object recognition is designed to effectively integrate information from three heterogeneous modalities: vision, touch, and text with the architecture illustrated in Fig. 1. The model adopts a feature-level fusion strategy. Initially, data from each modality undergoes independent feature extraction using dedicated networks. Patch Embedding is applied to visual and tactile inputs to capture spatial and structural features, while a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model processes textual data to extract semantic features. Parameter-sharing cross-attention mechanism then enables fine-grained cross-modal interactions. Specifically, visual features serve as the Query and interact with the Key/Value pairs from both tactile and textual modalities, producing visual-tactile and visual-text fusion features.

Similarly, tactile features as Query generate tactile-visual and tactile-text fusion features, and text features as Query yield text-visual and text-tactile fusion features. The two fused features corresponding to each original modality are then concatenated, forming enhanced unimodal representations that incorporate cross-modal information. Three enhanced unimodal features are then aggregated into a global fusion representation. This representation, augmented with a sequential dimension, is processed by a self-attention-based module to capture deeper cross-modal dependencies and improve feature discriminability.

Finally, the refined global features are passed through a classification head incorporating Layer Normalization and GELU activation functions to produce object category predictions. This architecture supports progressive feature learning from fine-grained cross-modal alignment to global semantic integration, maximizing the benefits of multimodal fusion while preserving the intrinsic properties of each modality. The following subsections provide detailed descriptions of the model components and implementation.

3.1 Feature Extraction

For the visual modality, the RGB input image $V \in \mathbb{R}^{3\times H\times W}$, where H and W represent the image height and width, is processed using a Patch Embedding layer, a technique commonly employed in Vision Transformers (ViT). The image is divided into non-overlapping patches of size $P\times P$. Each patch is then linearly projected into a high-dimensional feature vector of dimension C. These vectors are organized into an input sequence $V' \in \mathbb{R}^{N\times L\times C}$ for subsequent processing.Specifically, $L = (H/P) \times (W/P)$ represents the sequence length, C is the embedding dimension, and N indicates the batch size.

This transformation can be formally expressed as:

$$V' = \text{LinearProjection}(\text{Patchify}(V)) \tag{1}$$

where Patchify$(\cdot)$ splits the image into patches, and LinearProjection$(\cdot)$ maps each flattened patch to the embedding space C.

This approach treats each image patch as a visual token, preserving local spatial information within the sequence while facilitating subsequent attention-based processing.Similarly, for the tactile modality, RGB-formatted input data $T \in \mathbb{R}^{3\times H\times W}$ undergoes an identical patch embedding process. The tactile image is tokenized into

a sequence of embeddings $T' \in \mathbb{R}^{N \times L \times C}$ using the same patch size and projection dimension, ensuring structural alignment between the visual and tactile feature spaces.

Textual inputs in this study consist of natural language descriptions of object properties, provided by humans based on visual observation and tactile experience. Those manually annotated texts are processed using a pre-trained BERT model, which leverages the Transformer encoder architecture to extract rich semantic features. To align the textual features with the features from the visual and tactile modalities, a fully connected layer projects the 768-dimensional BERT embeddings into a target space of dimension C, matching the feature dimensions of the non-text modalities. This step effectively bridges the representational gap between natural language processing and computer vision, enabling cohesive multimodal integration.

Through the feature extraction module, inputs from visual, tactile, and textual modalities are transformed into aligned token sequences with consistent high-dimensional representations. Those sequences retain rich modality-specific information while meeting the input requirements of the subsequent cross-attention module.

3.2 Cross-attention Module

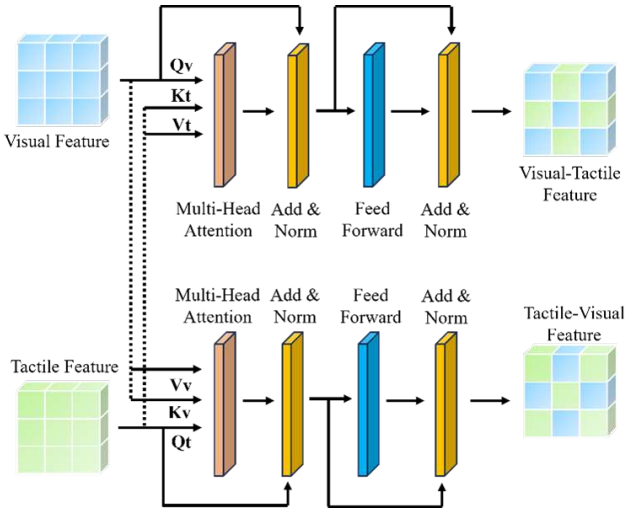


Fig. 2 Visual-Tactile Cross-Attention Operation. The bidirectional mechanism aligns features by using visual tokens to query tactile context (upper branch) and tactile tokens to query visual context (lower branch). Both streams process interactions through Multi-Head Attention and Feed-Forward layers to generate mutually enhanced representations.

The cross-attention module is designed to facilitate bidirectional interactions between each pair of the three input modalities. Building upon the standard self-attention mechanism, this module operates on the visual, tactile, and textual token sequences generated by the feature extraction module. The core cross-attention operation is formulated as follows:

$$Q = XW^Q, K = XW^K, V = XW^V \tag{2}$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where Q, K, and V form the core components of the attention mechanism. X represents the input token sequence, while W^Q, W^K, and W^V denote the corresponding weight matrices. The dot product of Q and K is used to compute the attention scores. To mitigate gradient instability during training, a scaling factor $\sqrt{d_k}$ is introduced, where d_k is the dimension of K. The softmax function normalizes the attention scores, which are used to compute a weighted sum over V.

In self-attention, Q, K, and V are all derived from the same input sequence, allowing the model to capture intra-modal dependencies. In contrast, cross-attention uses Q from one modality and K, V from another, allowing one modality to attend to and incorporate features from another. In the proposed framework, each modality serves as the Query sequence in turn, while the other two modalities provide the Key and Value sequences. For instance, when the visual tokens act as Q, and the tactile tokens as K and V, the output is a visual representation infused with tactile information. Similarly, when the visual tokens act as Q and the textual tokens as K and V, the result is a visual representation enriched with semantic attributes. The same process is applied to the tactile and textual modalities. Each modality thus generates two enhanced representations, each integrating information from one other modality. Those pairs are concatenated to form a final modality-specific representation that embodies cross-modal awareness. For instance, the visual output is a concatenation of visual-tactile and visual-textual features.

This design establishes six unique cross-modal interaction pathways (visual-tactile, visual-text, tactile-visual, tactile-text, text-visual, text-tactile), ensuring comprehensive information exchange across all three modalities. The cross-attention mechanism between visual and tactile is illustrated in Fig. 2, with similar operations applying to other modality pairs. The resulting token sequences not only retain their original modality characteristics but also embody coherent cross-modal correlations, providing semantically aligned inputs for subsequent fusion stages.

3.3 Self-attention Module

Following the cross-attention module, enhanced unimodal features are obtained, which incorporate complementary information from the other modalities. Those representations are then concatenated to form a unified multimodal representation that encapsulates the characteristics from all three modalities.

This combined feature sequence is processed by a self-attention module, which further captures complex dependencies and interactions among the three modalities.The self-attention operation refines the feature representation by modeling its internal structural relationships, thereby enhancing information fusion across modalities.The optimized global features are subsequently passed through a classification head, which consists of linear layers followed by a GELU activation function, to produce the final object category predictions.

To enhance training stability and generalization, Layer Normalization is applied for regularization using (4).

$$y = \frac{x - E(x)}{\sqrt{\mathrm{Var}(x) + \epsilon}}\gamma + \beta \tag{4}$$

where $E(x)$ and $\mathrm{Var}(x)$ represent the mean and variance of the features across all channels, ϵ is a small positive constant that prevents division by zero and avoids numerical instability, γ is a channel-specific learnable scale parameter, β is a channel-specific learnable shift parameter, and y is the final normalized output feature.This hierarchical approach effectively captures both fine-grained and high-level cross-modal interactions, resulting in discriminative multimodal representations that significantly enhance object recognition performance.

4 Dataset

Existing public datasets for robotic multimodal object recognition are often limited to specific material types and lack coverage of diverse everyday objects. Moreover, they frequently suffer from issues such as weak inter-modal correspondence and a scarcity of synchronized visual-tactile-textual triple data, making it difficult to adequately support the training and evaluation of multimodal fusion models in real-world interactive scenarios. Among existing datasets, the public multimodal dataset MSDO (Multi-Sense Daily Objects)[14] stands out as one of the few publicly available datasets that are relatively comprehensive and can support multimodal research involving visual, tactile, and textual data. It comprises data across 18 object categories, with each category including 100 visual images, 150 tactile images, and three textual adjectives.

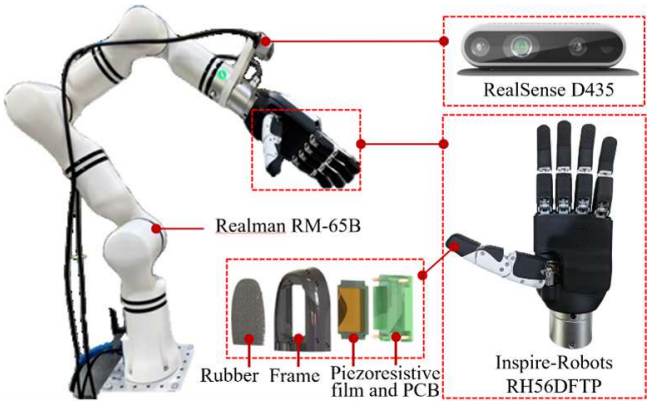


Fig. 3 Experimental Platform for Data Acquisition

To better validate model generalizability while advancing research in the field, this study constructed a new self-made dataset comprising visual, tactile, and textual data, acquired using a depth camera and a tactile dexterous hand as shown in Fig. 3.The dataset includes 20 categories of common daily objects, which vary not only

in visual appearance and tactile properties but also include instances with visual or tactile similarities.

Specifically, visual data were acquired using an Intel RealSense D435 depth camera, mounted to an Inspire-Robots tactile dexterous hand. In this study, only the RGB modality was utilized. Tactile data were collected using the same tactile dexterous hand, which is equipped with 1,062 densely distributed tactile sensing units from Wuhan Huaweike Intelligent Technology Co., Ltd. across its four fingers and palm. These sensors capture high-resolution contact information during physical interaction. The hand was installed on a collaborative robot arm (Realman RM-65B), and each object was subjected to a structured grasp-lift-reset-release procedure. Regarding the grasping strategy, side grasps were employed for cylindrical objects to ensure stable contact, while top-down grasps were used for all other object geometries.

For each object, approximately two hundred high-quality RGB images were captured from multiple viewpoints with a resolution of 640×480 pixels. These images provide clear visual features of the objects under standard lighting conditions. To ensure compatibility with the patch embedding layers in our network architecture, all images were preprocessed by resizing and center-cropping to a uniform resolution of 224×224 pixels.

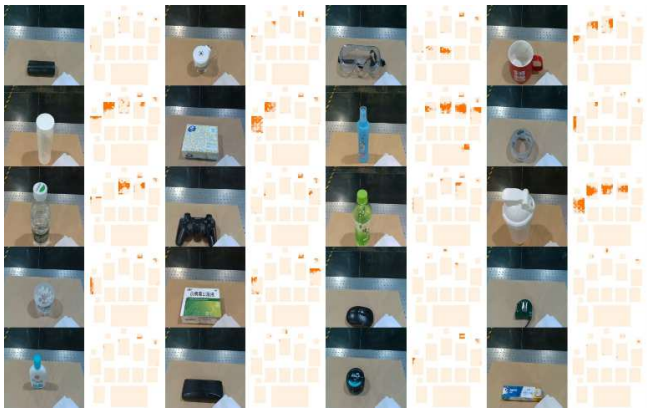


Fig. 4 Samples of Visual and Tactile Data

Tactile signals were sampled at 30 Hz, yielding approximately 160 tactile frames per object. Specifically, the raw pressure readings were normalized and spatially mapped onto a 2D canvas to reflect the physical layout of the tactile sensors.To adapt the single-channel tactile signals for the subsequent patch embedding layer—identical to the one used for the visual stream—we employed a pseudo-color mapping strategy.A custom linear segmented colormap was applied to transform scalar pressure values into three-channel RGB images, transitioning from light skin tone (low pressure) to deep orange (high pressure). Finally, these images were resized to 224×224 pixels to match the resolution of the visual data, ensuring that both modalities share a unified input dimension for consistent patch tokenization. Example tactile images and their visual counterparts are shown in Fig. 4.

Textual annotations were generated by human participants through natural language descriptions based on both visual observation and physical properties. To ensure consistency and objectivity, descriptors were restricted to visually observable attributes (e.g., color, shape) and tactilely perceptible properties (e.g., material, hardness). Each object was annotated using three adjectives, such as soft, cylindrical, and metallic.

Compared to the MSDO dataset, the self-collected dataset introduced in this study contains a greater number of object categories and provides a larger number of samples per modality. Those features contribute to enhanced robustness and generalizability in both model training and validation.

5 Experiment

To validate the effectiveness and robustness of the proposed approach, comprehensive experiments were conducted on both our self-collected dataset and the public MSDO dataset. All models were implemented in PyTorch 2.7.1 and trained on a high-performance workstation equipped with an Intel Core i9-14900K CPU and an NVIDIA RTX 4090 GPU. The training process utilized the Cross-Entropy Loss function and the AdamW optimizer, configured with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-2} to ensure stable convergence. Models were trained for 100 epochs with a batch size of 32.

Crucially, to assess the model's robustness against real-world visual perturbations, we adopted a rigorous evaluation protocol where the model was trained and tested on a dataset subjected to synthetic degradation. Specifically, the visual data underwent a combination of randomized quality distortions—simulating challenging conditions such as poor lighting, sensor noise, and partial occlusions—prior to training and evaluation (detailed degradation procedures are provided in the Appendix). This setup rigorously tests whether the integration of tactile and textual modalities can compensate for visual impairments when visual information is unreliable. Performance was monitored via loss curves and accuracy metrics, while Precision, Recall, and F1-score were computed from the confusion matrices to provide a holistic assessment of classification capability.

5.1 Experiment Results

The classification performance was evaluated under four input configurations: vision-only, tactile-only, vision–tactile fusion, and vision–tactile–text fusion.For the unimodal experiments, the cross-attention module was omitted; features were directly passed from the feature extraction module to the self-attention module and classification head. For the bimodal experiments, the original fusion pipeline was retained to accommodate the absence of the third modality.

Fig. 5 and Fig. 6 illustrate the model's classification performance on the self-collected and MSDO datasets, respectively. The loss curves for all configurations decreased consistently during training. Notably, the vision-tactile-text multimodal model converged faster than all others, demonstrating its strong learning capacity

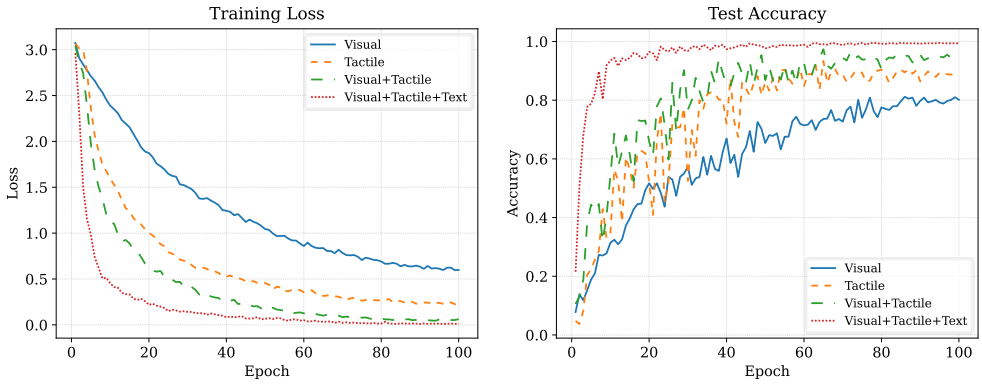


Fig. 5 Experimental Results on Self-Collected Dataset

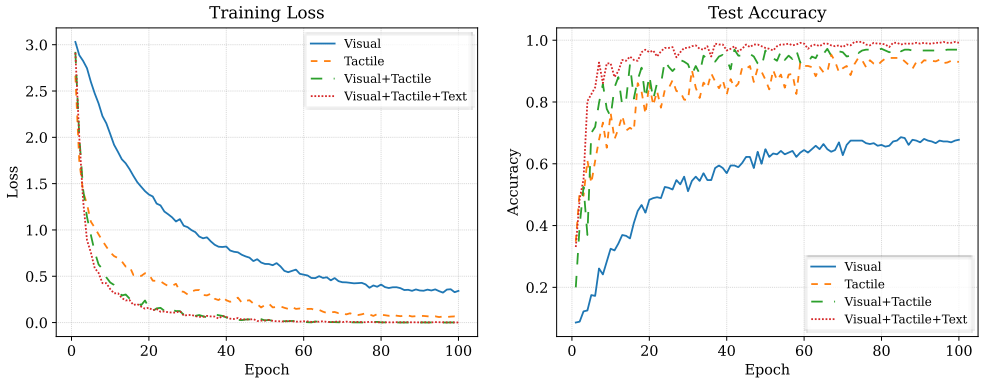


Fig. 6 Experimental Results on MSDO Dataset

and adaptability. The model achieved a recognition accuracy of 99.38% on the self-collected dataset and 99.17% on the MSDO dataset, indicating excellent performance in multimodal object recognition.

To further analyze the influence of different modalities, confusion matrices were generated for both datasets (as shown in Fig. 7 and Fig. 8), and standard performance metrics, including accuracy, recall, precision, and F1-score, were computed, as summarized in TABLE 1 and TABLE 2 and graphically illustrated in Fig. 9 and Fig. 10. Unimodal inputs achieved limited accuracy. For instance, vision-only recognition attained 67.78% on the MSDO dataset and 80.13% on the self-collected dataset; tactile-only recognition attained 88.85% on the self-collected dataset and 92.96% on the MSDO dataset. Bimodal fusion substantially improved performance, with the highest accuracy reaching 96.94% on the MSDO dataset. The proposed trimodal fusion strategy yielded the best overall performance, achieving a peak accuracy of 99.38% on the self-collected dataset.

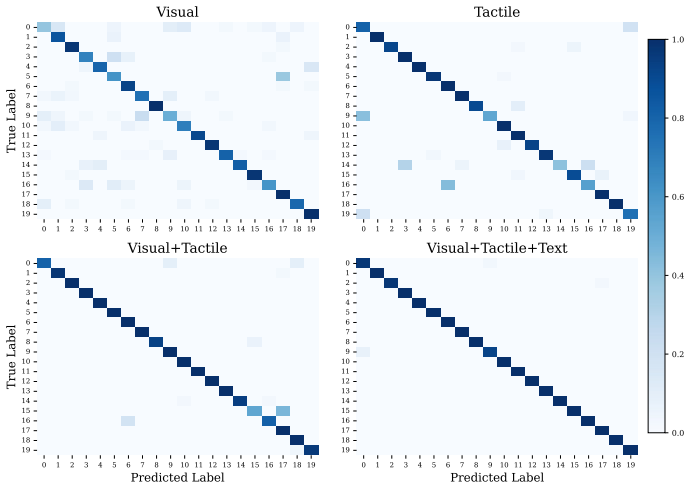


Fig. 7 Confusion Matrix of Self-Collected Dataset

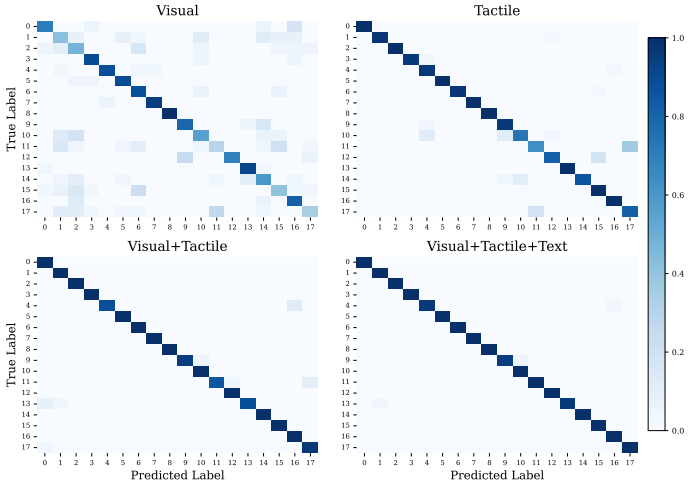


Fig. 8 Confusion Matrix of MSDO Dataset

Table 1 Performance Indicators of Self-collected Dataset

Modal	Accuracy	Recall	Precision	F1-score
Visual	80.13	80.45	80.14	79.50
Tactile	88.85	88.23	90.15	87.87
Visual+Tactile	95.20	94.82	95.90	94.85
Visual+Tactile+Text	**99.38**	99.32	99.32	99.31

Table 2 Performance Indicators of MSDO Dataset

Modal	Accuracy	Recall	Precision	F1-score
Visual	67.78	69.54	69.75	68.50
Tactile	92.96	92.65	93.10	92.61
Visual+Tactile	96.94	97.33	97.02	97.01
Visual+Tactile+Text	**99.17**	99.26	99.14	99.18

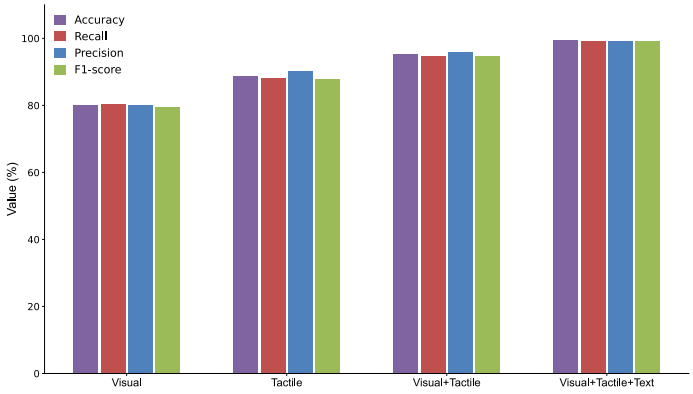


Fig. 9 Comparison Histograms of Self-collected Dataset

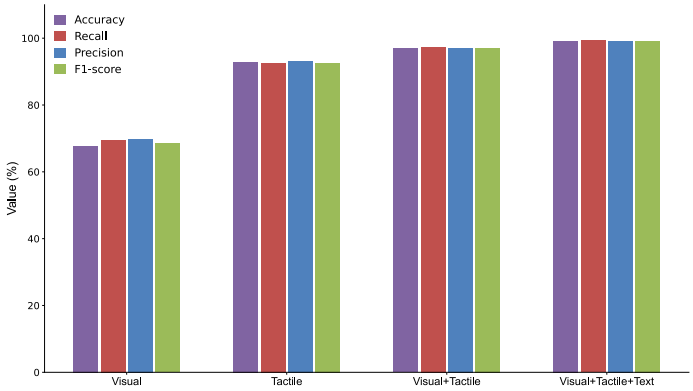


Fig. 10 Comparison Histograms of MSDO Dataset

To validate the effectiveness of the proposed cross-attention fusion module, we conducted ablation studies by comparing our model with a baseline that employs simple feature concatenation followed by self-attention. The experimental results showed in TABLE 3 demonstrate the superiority of our approach: on the self-collected dataset, the proposed model achieved an accuracy of 99.38%, outperforming the baseline variant which reached 97.23%; similarly, on the MSDO dataset, our model attained an accuracy of 99.17%, compared to 97.33% for the baseline. These findings confirm that

the cross-attention mechanism facilitates more effective and fine-grained inter-modal interaction than direct concatenation.

Table 3 Ablation Results with Both Datasets

Methods	Dataset	Accuracy (%)
w/o Cross-Attention	Self-made	97.23
	MSDO	97.33
Ours	Self-made	**99.38**
	MSDO	**99.17**

Given the prevalence of self-made datasets in multimodal fusion research, direct performance comparisons across studies are often challenging due to variations in data sources and experimental setups. TABLE 4 summarizes representative recent works in this field, listing their models, datasets, and reported accuracy to provide a comparative overview of the state-of-the-art. While a direct benchmark is difficult due to the lack of open-source code and differing task objectives in some studies, this comparison highlights the efficacy of our proposed method. Our model demonstrates robust performance on both our self-collected dataset and the publicly available MSDO dataset, successfully integrating visual, tactile, and textual semantics. Notably, compared to the current SOTA model VITO on similar multimodal tasks, our proposed Fusion-VTT achieves superior results in both object recognition precision and overall accuracy, validating the effectiveness of our fusion architecture.

Table 4 Comparative Results with Other Methods

Methods	Dataset	Model	Accuracy (%)
DGML [31]	Self-made	CNN+LSTM	93.10
TacNet [32]	Self-made	CNN	95.02
VITO [33]	Self-made	Transformer	95.60
	ObjectFolder2.0		97.40
Ours	Self-made	Fusion-VTT	**99.38**
	MSDO		**99.17**

5.2 Discussion and Analysis

The above experiments validate the effectiveness of the proposed two-stage fusion framework in processing multimodal inputs. Ablation studies and comparative tests confirm its advantages over alternative fusion methods.

Multimodal fusion substantially enhanced recognition accuracy. The proposed model achieved accuracies of 99.17% and 99.38% on both datasets, respectively. These results demonstrate not only the effectiveness of feature-level fusion in processing heterogeneous data, but also the robustness of the cross-modal complementary mechanism

under complex scenarios. Ablation experiments further highlighted the contribution of each modality: the removal of textual information resulted in a 4.18% decrease in accuracy, underscoring the importance of semantic cues in resolving material ambiguities.

Modal imbalance represents a critical challenge in multimodal learning systems. Conventional fusion approaches often depend on manually tuned weighting coefficients, which limit their adaptability and generalizability. In contrast, the proposed hierarchical fusion framework balances the preservation of modality-specific characteristics with the enhancement of cross-modal interactions. During the first fusion stage, each modality generates independent query, key, and value vectors, enabling bidirectional interactions through a parameter-shared cross-attention mechanism. The second stage utilizes a self-attention layer to dynamically integrate features into a globally coherent representation. Compared to conventional fusion strategies, which rely on simple feature concatenation, the proposed design enables the model to autonomously learn inter-modal relationships during training, thereby effectively mitigating the issue of modal imbalance.

6 Conclusion

A novel hierarchical fusion framework for multimodal object recognition is introduced in this paper, which achieves deep integration of visual, tactile, and textual modalities through collaborative perception. The proposed two-stage fusion strategy effectively addresses the challenges of aligning and modeling interactions among heterogeneous modalities, significantly enhancing recognition performance in complex scenarios. A new multimodal dataset comprising 20 categories of common daily objects is constructed, addressing the scarcity of synchronized visual-tactile-textual data for object recognition. Experimental validation demonstrates that the proposed approach achieves recognition accuracies of 99.38% and 99.17% on the self-collected and MSDO datasets, respectively, outperforming the best baseline method by 1.98%. These results underscore the substantial potential of multimodal fusion and highlight the value of the newly constructed dataset in validating methodological efficacy.

While the proposed method yields considerable performance improvements, several directions warrant further investigation. Future work will incorporate dynamic tactile sequence modeling to leverage temporal information, explore unsupervised modal alignment techniques to reduce annotation dependency, and expand the diversity of object categories and environmental contexts. The integration of fine-tuning strategies will also be investigated to enhance deployment in resource-constrained platforms. These developments are expected to further advance the perceptual capabilities of robotic systems in complex environments.

Declarations

Author contribution Investigation, methodology, experiments, writing—original draft preparation, Yang Li; methodology, validation, writing—review and editing, Liangqi Zhang and Min Li. All authors have read and agreed to the published version of the manuscript.

Funding This work was supported in part by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China under Grant JYB2025XDXM208, the National Natural Science Foundation of China under Grants 52575019 and 52188102, and the National Natural Science Foundation of China Overseas Excellent Young Scholars Fund.

Competing interests The authors declare no competing interests.

Appendix A Visual Degradation Protocol

To simulate challenging real-world visual conditions such as poor lighting, sensor noise, and partial occlusions, a randomized degradation pipeline was applied to the original high-quality RGB images. This process generates the degraded visual dataset used for training and evaluation in Section 5. The specific degradation operations and their parameter ranges are detailed below:

- **Brightness Adjustment:** Randomly scales the image brightness factor between [0.3, 1.4], simulating environments ranging from severe low-light to overexposure.
- **Contrast Scaling:** Randomly adjusts the contrast factor within [0.6, 1.4] to mimic varying visibility conditions.
- **Gaussian Noise Injection:** Adds Gaussian noise with a standard deviation randomly sampled from [5, 40] to 70% of the images, simulating sensor grain.
- **Defocus Blur:** Applies a Gaussian blur with a kernel size randomly chosen from {1, 3, 5, 7, 9} to 50% of the images, simulating out-of-focus capture.
- **Shadow Simulation:** Overlays a linear gradient shadow mask with an intensity between [0.2, 0.6] to 60% of the images, replicating uneven lighting or partial shadows.

Fig. A1 illustrates representative examples of these degraded samples compared to the original high-quality inputs.

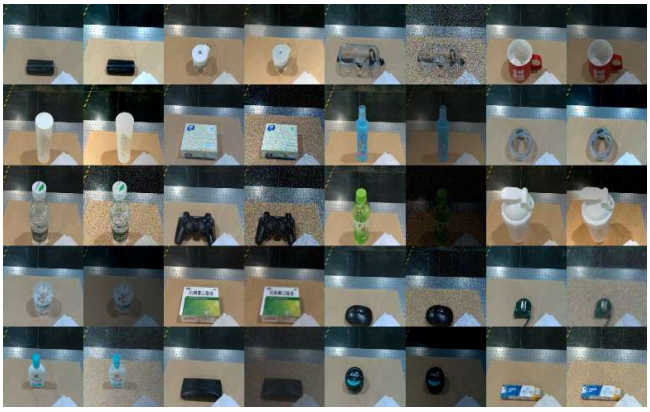


Fig. A1 Samples of Degraded and Original Visual Data

Appendix B Mapping of Object Labels

In the confusion matrices presented in Section 5, numeric indices are used to represent object categories to ensure visualization clarity and compactness. TABLE B1 provides the complete correspondence between these indices and the actual object category names for both datasets.

Table B1 Mapping between numeric indices and real object categories

Index	Self-collected Dataset	MSDO Dataset
0	bin-liner	adhesive
1	bottle	allen
2	cestbon	ball
3	cotton swab	bottle
4	dabao	box
5	dianwenx	branch
6	drawer	caliper
7	gamepad	can
8	ganmaol	hand
9	glasses case	highlighter
10	goggles	key
11	hualushui	pen
12	lvcha	plier
13	mouse	rock
14	mud	rubber
15	mug	scissors
16	scotch tape	sticky tape
17	shake cup	tube
18	tapeline	N/A
19	toothpaste	N/A

References

[1] Lu, K., An, X., Li, J., He, H.: Efficient deep network for vision-based object detection in robotic applications. Neurocomputing **245**, 31–45 (2017)

[2] Luan, C., Cui, Z. Ruyao Lin, Sun, L.: A siamese network utilizing image structural differences for cross-category defect detection. In: 2020 IEEE International Conference on Image Processing (ICIP), pp. 778–782 (2020)

[3] Russakovsky, O., Deng, J., Su, H., Krause, J., *et al.*: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)

[4] Petriu, E.M., Payeur, P., Cretu, A.-M., Pasca, C.: Complementary tactile sensor and human interface for robotic telemanipulation. In: 2009 IEEE International Workshop on Haptic Audio Visual Environments and Games, pp. 164–169 (2009)

[5] Yuan, W., Dong, S., Adelson, E.H.: Gelsight: High-resolution robot tactile sensors for estimating geometry and force. Sensors **17**(12), 2762 (2017)

[6] Kerr, E., McGinnity, T.M., Coleman, S.: Material recognition using tactile sensing. Expert Systems with Applications **94**, 94–111 (2018)

[7] Yuan, W., Srinivasan, M.A., Adelson, E.H.: Estimating object hardness with a gelsight touch sensor. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 208–215 (2016)

[8] Gandarias, J.M., Gómez-de-Gabriel, J.M., García-Cerezo, A.J.: Enhancing perception with tactile object recognition in adaptive grippers for human–robot interaction. Sensors **18**(3), 692 (2018)

[9] Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E., Hussain, A.: Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. Information Fusion **91**, 424–444 (2023)

[10] Babadian, R.P., Faez, K., Amiri, M., Falotico, E.: Fusion of tactile and visual information in deep learning models for object recognition. Information Fusion **92**, 313–325 (2023)

[11] Li, S., Wang, R., Zheng, P., Wang, L.: Towards proactive human–robot collaboration: A foreseeable cognitive manufacturing paradigm. Journal of Manufacturing Systems **60**, 547–552 (2021)

[12] Cheng, C., Xu, T., Wu, X.-J., Li, H., *et al.*: Textfusion: Unveiling the power of textual semantics for controllable image fusion. Information Fusion **117**, 102790 (2025)

[13] Chen, Y., Sipos, A., Merwe, M.V., Fazeli, N.: Visuo-tactile transformers for manipulation. arXiv preprint arXiv:2210.00121 (2022)

[14] Li, B., Li, L., Wang, H., Chen, G., *et al.*: Tvt-transformer: A tactile-visual-textual fusion network for object recognition. Information Fusion **118**, 102943 (2025)

[15] Yuan, W., Wang, S., Dong, S., Adelson, E.H.: Connecting look and feel: Associating the visual and tactile properties of physical materials. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4494–4502 (2017)

[16] Strese, M., Schuwerk, C., Steinbach, E.: Surface classification using acceleration signals recorded during human freehand movement. In: 2015 IEEE World Haptics Conference (WHC), pp. 214–219 (2015)

[17] Luo, S., Yuan, W., Adelson, E.H., Cohn, A.G., Fuentes, R.: Vitac: Feature sharing

between vision and tactile sensing for cloth texture recognition. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 2722–2727 (2018)

[18] He, B., Wang, Z., Zhou, Y., Chen, J., *et al.*: Microsaccade-inspired event camera for robotics. Science Robotics **9**(90), 8124 (2024)

[19] Corradi, T., Hall, P., Iravani, P.: Object recognition combining vision and touch. Robotics and Biomimetics **4**(1), 2 (2017)

[20] Nisal, K., Ruhunge, I., Subodha, J., Perera, C.J., Lalitharatne, T.D.: Design, implementation and performance validation of uompro artificial hand: Towards affordable hand prostheses. In: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 909–912 (2017)

[21] James, J.W., Pestell, N., Lepora, N.F.: Slip detection with a biomimetic tactile sensor. IEEE Robotics and Automation Letters **3**(4), 3340–3346 (2018)

[22] Wang, D., Teng, Y., Peng, J., Zhao, J., *et al.*: Deep-learning-based object classification of tactile robot hand for smart factory. Applied Intelligence **53**, 22374–22390 (2023)

[23] Yuan, Q., Wang, J.: Design and experiment of the nao humanoid robot's plantar tactile sensor for surface classification. In: 2017 4th International Conference on Information Science and Control Engineering (ICISCE), pp. 931–935 (2017)

[24] Gao, J., Li, P., Chen, Z., Zhang, J.: A survey on deep learning for multimodal data fusion. Neural Computation **32**(5), 829–864 (2020)

[25] Takahashi, K., Tan, J.: Deep visuo-tactile learning: Estimation of tactile properties from images. In: 2019 International Conference on Robotics and Automation (ICRA), pp. 8951–8957 (2019)

[26] Zhang, T., Cong, Y., Dong, J., Hou, D.: Partial visual-tactile fused learning for robotic object recognition. IEEE Transactions on Systems, Man, and Cybernetics: Systems **52**(7), 4349–4361 (2022)

[27] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., *et al.*: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML), pp. 8748–8763 (2021)

[28] Huang, Z., Lv, C., Xing, Y., Wu, J.: Multi-modal sensor fusion-based deep neural network for end-to-end autonomous driving with scene understanding. IEEE Sensors Journal **21**(10), 11781–11790 (2021)

[29] Lin, Z., Yu, S., Kuang, Z., Pathak, D., Ramanan, D.: Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In: 2023

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 19325–19337 (2023)

[30] Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., *et al.*: Imagebind: One embedding space to bind them all. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15180–15190 (2023)

[31] Anzai, T., Takahashi, K.: Deep gated multi-modal learning: In-hand object pose changes estimation using tactile and image data. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 9361–9368 (2020)

[32] Gandarias, J.M., Garcia-Cerezo, A.J., Gomez-de-Gabriel, J.M.: Cnn-based methods for object recognition with high-resolution tactile sensors. IEEE Sensors Journal **19**(16), 6872–6882 (2019)

[33] Li, B., Bai, J., Qiu, S., Wang, H., Guo, Y.: Vito-transformer: A visual-tactile fusion network for object recognition. IEEE Transactions on Instrumentation and Measurement **72**, 1–10 (2023)