RAV: Retrieval-Augmented Voting for Tactile Descriptions Without Training

Jinlin Wang[1], Hongyu Yang[2,*], Yulong Ji[3,*]
[1]National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu, 610065, China.
[2]College of Computer Science, Sichuan University, Chengdu, 610065, China.
[3]School of Aeronautics and Astronautics, Sichuan University, Chengdu, 610065, China.
Correspondence: jyl@scu.edu.cn , yanghongyu@scu.edu.cn

Abstract

Tactile perception is essential for human-environment interaction, and deriving tactile descriptions from multimodal data is a key challenge for embodied intelligence to understand human perception. Conventional approaches relying on extensive parameter learning for multimodal perception are rigid and computationally inefficient. To address this, we introduce Retrieval-Augmented Voting (RAV), a parameter-free method that constructs visual-tactile cross-modal knowledge directly. RAV retrieves similar visual-tactile data for given visual and tactile inputs and generates tactile descriptions through a voting mechanism. In experiments, we applied three voting strategies, SyncVote, DualVote and WeightVote, achieving performance comparable to large-scale cross-modal models without training. Comparative experiments across datasets of varying quality—defined by annotation accuracy and data diversity—demonstrate that RAV's performance improves with higher-quality data at no additional computational cost. Code, and model checkpoints are opensourced at https://github.com/PluteW/RAV.

1 Introduction

Biological perception is inherently multimodal, with tactile perception enabling humans to discern object properties like shape, texture, and hardness through environmental interaction (Navarro-Guerrero et al., 2023; Zhong et al., 2024). In robotics, tactile perception is vital for task generalization and adaptation (Bonner et al., 2021; Wang, 2024). Yet, compared to vision and audition, tactile perception remains underexplored. Current research primarily integrates vision and touch for action execution (Han et al., 2021; Qi et al., 2023) object categorization (Cheng et al., 2024) or materials categorization (Cheng et al., 2024), relying on large-scale datasets to train complex models, which face diminishing returns under scaling laws (Dettmers and Zettlemoyer, 2023). Notably, knowledge-driven approaches for visual-tactile cross-modal tasks are limited, representing a significant research gap.

To address this, we propose Retrieval-Augmented Voting (RAV), a parameter-free method that enhances cross-modal perception via a visual-tactile knowledge base. Using CLIP (Radford et al., 2021), we construct separate vector databases for visual and tactile features. In order to collect the knowledge and obtain the final output, we design the mechanism of voting. For input visual-tactile data, RAV retrieves similar features and generates tactile descriptions through voting, employing three strategies: SyncVote (equal voting), DualVote (sensory credibility-based voting), and WeightVote (distance-weighted voting). Unlike parameter-intensive models, RAV efficiently leverages existing data, inspired by retrieval-augmented generation in language models (Lewis et al., 2020; Jiang et al., 2024). Experiments on tactile description tasks show that RAV achieves performance comparable to large-scale trained models on the same dataset. Additionally, the method demonstrates the ability to enhance performance as data quality improves, incurring minimal additional cost, which represents a significant advantage over parameter-dependent models.

The main contributions of the work can be described as follows:

1. Proposal of RAV, a parameter-free visual-tactile cross-modal perception method, using CLIP to construct a knowledge base and employing three voting strategies (SyncVote, DualVote, WeightVote) for tactile description generation;
2. Achievement of performance comparable to large cross-modal models in tactile description tasks, overcoming parameter training lim-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 6199–6206
November 4-9, 2025

itations;

3. Demonstration of low computational cost and performance scalability with improved data quality;
4. Validation of RAV's effectiveness in leveraging existing data, highlighting visual-tactile cross-modal perception challenges.

2 Realted work

2.1 Tactile Perception

The exploration of tactile perception is a critical avenue for advancing the development of robots towards greater generalization and adaptability. The majority of tactile perception data is acquired through tactile sensors, including vision-based tactile sensors, such as GelSight (Yang et al., 2022), DIGHT (Kerr et al., 2023), and GelSlim (Gao et al., 2023). These sensors are capable of detecting changes in surface texture during contact by a micro-camera underneath the elastic gel.

With the development of these advanced tactile sensors, several publicly available high-quality tactile datasets have been obtained, including TVL (Fu et al., 2024), SSVTP (Kerr et al., 2023), Touch and Go (Yang et al., 2022), Touch100k (Cheng et al., 2024), support tasks like material classification and tactile description. However, existing methods rely on large-scale parameter training, limiting efficiency. For instance, Yang et al. (Yang et al., 2022) obtained 54.7% material classification accuracy and 77.3% material attribute recognition accuracy by 240 epochs of comparative learning on both visual and tactile images; while Fu et al. (Fu et al., 2024) attained 81.7% classification accuracy and strong tactile description performance by adding 86M parameters to LLaMa2. Such approaches require complex training, struggle to leverage existing data efficiently, and lack knowledge-driven mechanisms for cross-modal perception, hindering further advancements.

2.2 Retrieval-Augmented Generation (RAG)

Recent years have witnessed the surge of interest in Artificial Intelligence Generated Content (AIGC). Despite significant progress in generative modeling, AIGC still faces challenges such as outdated knowledge and lack of long-tail knowledge (Mallen et al., 2023). An effective solution to such problems is through retrieval. The purpose of retrieval is to identify relevant existing objects from a large number of resources. Efficient information retrieval systems can handle document collections of billions of orders of magnitude (Johnson et al., 2021). In addition to documents, retrieval is applied to many other modalities (Wu et al., 2024; Liu et al., 2024).

In RAG processing, given an input query, the retriever identifies relevant data sources and the retrieved information interacts with the generator to improve the generation process. While the concept of RAG originally emerged in the context of text-to-text generation processes, the technique has also been used in a variety of domains, including code (Parvez et al., 2021), audio (Koizumi et al., 2020; Huang et al., 2023), image (Sarto et al., 2022). The basic ideas and processes of RAG are largely consistent across the various paradigms. However, in the field of visual-tactile perception, there is still a gap in research directed at the retrieval and utilization of knowledge.

3 Methods

This study introduces Retrieval-Augmented Voting (RAV), a parameter-free visual-tactile cross-modal perception method that constructs a knowledge base using CLIP to extract features from visual and tactile images, stores them in a vector database, and fuses multimodal information via voting to generate tactile descriptions. As shown in Fig.1, RAV comprises multimodal retrieval and voting modules for retrieving similar features and weighted decision-making, respectively. Without any trainable parameters, RAV achieves high performance in tactile description tasks. Each module is detailed below.

3.1 Preliminary

The input data is formally denoted as X, and the corresponding features are represented as x. It is assumed that objects with similar feature vectors correspond to inputs that are also close. For any input, by comparing the feature similarity with cosine similarity:

$$sim(x, \dot{x}) = \frac{x \cdot \dot{x}}{\|x\|\|\dot{x}\|}, \tag{1}$$

where $\dot{x}$ denotes any instance in space $\mathbb{R}$, we are able to find the k closest instances, each with j_k labels, and model could output a list of the labels L after each instance votes. For visual and tactile inputs, we perform the above retrieval process separately and summarize the output in the voting stage.

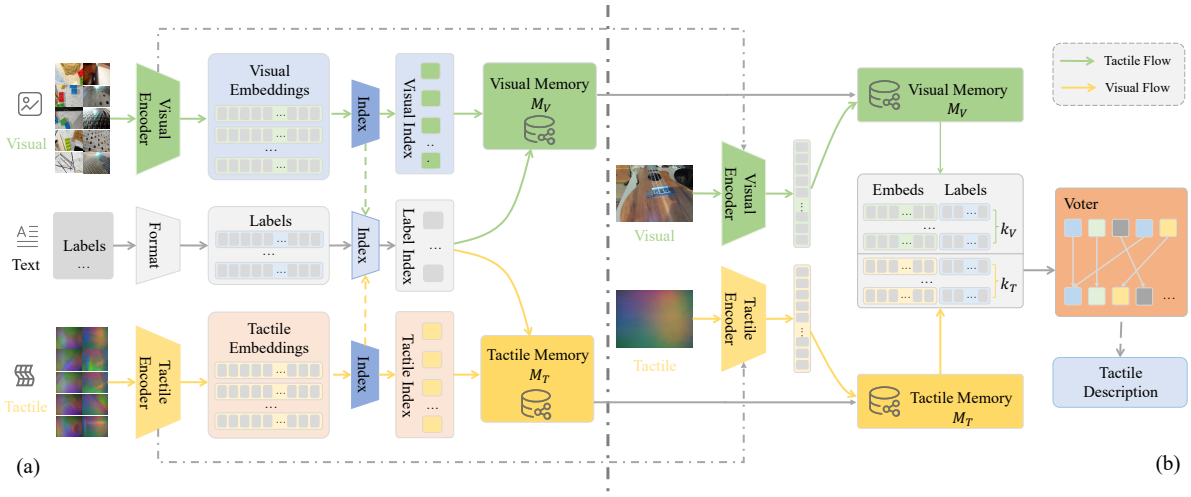

Figure 1: (a) Memory construction. Through the encoder, visual and tactile image data are separately processed to extract feature vectors. These feature vectors are paired with semantic labels and stored in the visual vector database $\mathcal{M}_\mathcal{V}$ and tactile vector database $\mathcal{M}_\mathcal{T}$, respectively.
(b) Query Process. Upon receiving visual and tactile image inputs, the same encoder extracts feature vectors. Relevant data are then retrieved from $\mathcal{M}_\mathcal{V}$ and $\mathcal{M}_\mathcal{T}$ based on vector distance, yielding visual and tactile label sets L_v and L_t. These multimodal label sets are input to the voter, which fuses them using SyncVote, DualVote, or WeightVote strategies to produce the final tactile description labels.

The inputs of visual and tactile are denoted as V_i and T_i. The corresponding features are denoted as v_i, t_i, respectively. The k_v instances most similar to v_i are identified through a search of the visual database, with the label set of each instance represented as $L_v = \{l_1, l_2, \ldots, l_n\}$. The k_t instances most similar to t_i from tactile database has labels denoted as $L_t = \left\{l'_1, l'_2, \ldots, l'_m\right\}$. After that, voter will output the list of cross-modal labels:

$$L = Vote(L_v, L_t). \tag{2}$$

In this way, the knowledge related to input x is extracted in the form of label list L.

3.2 Multimodal Retriever

The multimodal retriever queries vector databases $\mathcal{M}_\mathcal{V}$ and $\mathcal{M}_\mathcal{T}$ to retrieve instances similar to input visual and tactile features. The main challenge in the multimodal retrieval process is to efficiently encode and store a large number of visual / tactile / textual embeddings for fast and accurate retrieval.

Given a dataset D containing data samples (V_i, T_i, L_i) of a visual image V_i, a tactile image T_i and labels L_i, the CLIP image editor $\phi_i mg$ is used to extract the visual embedding $v_i \in \mathbb{R}^{d_v}$ and the tactile embedding $t_i \in \mathbb{R}^{d_t}$, respectively. The symbol d_v and d_t refers to the feature dimensions. The visual and tactile embeddings are stored in the memories $\mathcal{M}_\mathcal{V}$ and $\mathcal{M}_\mathcal{T}$ with lables L_i, as shown in Fig.1 (a).

3.3 Voter

After successfully constructing visual memory $\mathcal{M}_\mathcal{V}$ and tactile memory $\mathcal{M}_\mathcal{T}$ using the multimodal retriever, next step is to combine the memories with the retrieval process by integrating the retrieval results using a voter to improve the performance of the task. For input $X_i = \{V_i, T_i\}$, the CLIP model ϕ extracts visual features $v_i \in \mathbb{R}^{d_v}$ and tactile features $t_i \in \mathbb{R}^{d_t}$.

The embeddings v_i and t_i will then be navigated through the previously constructed indexes, and sorted according to similarity, the memory $\mathcal{M}_\mathcal{V}$ produces a list of the retrieved top k_v items with lables denoted as $L_v = \{L_1, L_2, \ldots, L_{k_v}\}$. Memory $\mathcal{M}_\mathcal{T}$ produces a list of the top k_t items with lables denoted as $L_t = \{L_1, L_2, \ldots, L_{k_t}\}$. In our experiments, we set k_v and k_t to 5.

The voter fuses L_v and L_t, outputting a cross-modal label list $L = Vote(L_v, L_t)$ for tactile description. Three voting strategies are designed here:

SyncVote. A balanced voting strategy assigns equal voting weights to labels from instances retrieved from visual and tactile vector databases $\mathcal{M}_\mathcal{V}$ and $\mathcal{M}_\mathcal{T}$. In this approach, each instance contributes one unit vote per label, assuming equal credibility for visual and tactile labels. Votes are aggregated via simple counting to fuse visual label set

L_v and tactile label set L_t, yielding a cross-modal label list L.

DualVote. An unbalanced voting strategy assigns weights based on the sensory origin of labels, determined by experimental observations and sensory specificity. Vision-related labels (e.g., "matte" or "smeared") from the visual vector database $\mathcal{M}_\mathcal{V}$ are deemed more credible and assigned higher weights (e.g., 1.5) than equivalent labels from the tactile vector database $\mathcal{M}_\mathcal{T}$ (e.g., 0.4). Conversely, tactile-related labels (e.g., "undulating" and "sticky") from $\mathcal{M}_\mathcal{T}$ are considered more credible and receive higher weights (e.g., 1.2) than those from $\mathcal{M}_\mathcal{V}$ (e.g., 0.6). Labels not clearly categorized are assigned unit weight. Label categorization is supported by GPT-4V's vision-language model, full categorization in the code.

WeightVote. An unbalanced voting strategy assigns weights based on the vector distance of retrieved instances to enhance cross-modal label fusion in tactile description tasks. It is assumed that labels from smaller vector distance instances possess greater credibility, more accurately reflecting the input object's properties. Denoting vector distance as z, the label weight is computed via the sigmoid function:

$$\sigma(z) = \frac{1}{1 + e^{-a(b-z)}}, \tag{3}$$

where hyperparameters a and b control the steepness and offset of the weight curve, ensuring higher weights for labels from smaller vector distance instances. Weights are applied to instances retrieved from visual and tactile vector databases $\mathcal{M}_\mathcal{V}$ and $\mathcal{M}_\mathcal{T}$.

4 Experiment and Result

This section quantitatively evaluated the capabilities of the RAV model in the tactile-semantic description task.

4.1 Dataset

Experiments utilize the TVL dataset (Fu et al., 2024), a cross-modal dataset comprising 44,000 visual-tactile pairs. TVL integrates two subsets: 1) the Supervised Visuo-Tactile Pretraining (SSVTP) dataset (Kerr et al., 2022), with 4,587 pairs collected by a UR5 robot capturing top-down visual images followed by vertical DIGIT sensor presses for tactile images; 2) the Human Collected Tactile (HCT) dataset (Fu et al., 2024), with instances of visual-tactile data gathered by five individuals in non-laboratory settings using a handheld device at 30 Hz, recording trajectories of approach, contact, sliding, and withdrawal, with visual data captured at an oblique angle to keep tactile sensors in view. Since the HCT dataset includes the entire process of acquisition, there are a large number of non-contact data samples.

TVL pairs are annotated with textual labels: 10% of the SSVTP subset (4,587 pairs) is manually labeled in English using a 400-word tactile vocabulary (ajbarnett, 2024) to describe material properties and tactile sensations, while the remaining 90% are pseudo-labeled by GPT-4V. The test set (1%) is manually annotated to ensure evaluation reliability.

Table 1: The voting weight of a label from the source and has the different type. The categorization of the labels comes from the GPT-4V and the weights used come from human experience and have not been fine-tuned.

Source	Type	Weight
Vision	Vision	1.5
	Tactile	0.6
	Unclear	1
Tactile	Vision	0.4
	Tactile	1.2
	Unclear	1

4.2 Tactile Description

In the tactile description task, the model receives visual and tactile image pairs as input and outputs a linguistic description of material properties and tactile sensations, which is limited to a maximum of five words. In the testing phase, to obtain numerical comparison results, we follow Letian's method (Fu et al., 2024) by scoring the similarity of output to the real labels through the GPT-4V on a scale of 1 to 10 (higher scores indicate that the model's outputs are closer to the human descriptions). Additionally, we ask GPT-4V to provide an interpretation of the corresponding scores, similar to prior works (Liu et al., 2023a,b).

For the DualVote strategy, weight parameters were determined via grid search over the range [0.2, 2.0] with a step size of 0.1, The search aimed to optimize accuracy on the TVL validation set, yielding weights as shown in Tab.1. For the WeightVote strategy, parameter a controls the steepness of the weight curve, and b adjusts the vector distance

Table 2: Tactile Description Score Sheet. RAV methods were compared with large models of multiple sizes on three datasets. Most methods pre-train the encoder on both visual and language tasks. Blue indicates the best performing result among the results requiring training, and red indicates the best performing method.

Model	Encoder Pre-training			Paramter Size	Score(1-10)		
	Vision	Tactile	Language		SSVTP	HCT	TVL
BLIP-2	✓	✗	✓	6.7 B	2.02	2.72	2.64
LLaVA-1.5(7B)	✓	✗	✓	7 B	3.64	3.55	3.56
ViP-LLaVA(7B)	✓	✗	✓	7 B	2.72	3.44	3.36
LLaMA-Adapter	✓	✗	✓	7 B	2.56	3.08	3.02
InstructBLIP(7B)	✓	✗	✓	7 B	1.40	1.30	1.31
SSVTP-LLaMA	✓	✓	✗	7 B	2.58	3.67	3.54
TVL-LLaMA	✓	✓	✓	7 B	6.16	4.89	5.03
LLaVA-1.5(13B)	✓	✗	✓	13 B	3.55	3.63	3.62
ViP-LLaVA(13B)	✓	✗	✓	13 B	4.10	3.76	3.80
InstructBLIP(13B)	✓	✗	✓	13 B	1.44	1.21	1.24
GPT-4V	✓	✗	✓	-	5.02	4.42	4.49
Clip-KNN-Vote	✓	✗	✗	0	5.47	4.58	4.83
RAV(SyncVote)	✓	✗	✗	0	6.18	4.88	5.08
RAV(DualVote)	✓	✗	✗	0	5.96	4.83	5.01
RAV(WeightVote)	✓	✗	✗	0	5.13	4.93	4.99

threshold, prioritizing labels from high-similarity instances. Based on the TVL set's vector distance distribution, $a = 0.4$ ensures a smooth weight transition, avoiding excessive penalties for low-similarity labels; $b = 9$ aligns the curve's center with the distance distribution range.

Summary statistics of the tactile description output results are provided in Tab.2. Although most methods employ encoders that are pretrained on both visual and language tasks, the open-source vision language models (VLMs) do not perform as well as GPT-4V on the benchmarks. This is attributed to the limited diversity of visual data utilized in their training and the lack of emphasis on human tactile sensations. For a more direct comparison of parameter-free methods, we also evaluated a strong baseline, Clip-KNN-Vote, which uses a simple majority vote on retrieved neighbors. While Clip-KNN-Vote achieves commendable performance, our RAV methods consistently outperform it, demonstrating that RAV's voting strategies are key to fuse effectively. On the other hand, the results of RAVs trained without any parameters were able to significantly outperform GPT-4V, while slightly outperforming the optimal generative model TVL. This suggests that our knowledge-based approach is able to obtain good cross-modal perception without any trainable parameters.

With the exception of TVL-LLaMA, which is designed for the task, and GPT-4V, which contains a large number of parameters, a wide range of methods involved in the comparison scored lower. Notably, InstructBLIP(13B) does not outperform its 7B counterpart and lags behind models of similar scale, suggesting that increasing parameter count alone has limited impact on enhancing cross-modal perception.

To further quantify the linguistic quality of the generated descriptions, we conducted an evaluation using standard NLP metrics, including BLEU, CIDEr, METEOR, and ROUGE. [1] The results showed that BLEU and CIDEr scores were consistently zero across all models. This is an expected outcome, as the tactile description task involves generating a few keywords rather than complete sentences, leading to sparse n-gram overlap that renders these metrics unsuitable for this application.

In contrast, METEOR and ROUGE proved to be more robust for this lexical-level task, as their evaluation mechanisms account for synonyms, stemming, and longest common subsequences. As presented in Tab.3, the RAV(SyncVote) strategy consistently achieved the highest scores on both METEOR and ROUGE across all three datasets. It

[1] All metrics were implemented using the nltk, pycocoevalcap, and rouge_score libraries with default parameter settings.

Table 3: Tactile Description NLP Metrics Evaluation. RAV methods were compared with large models of multiple sizes on three datasets. Red indicates the best performing method.

Model	SSVTP		HCT		TVL	
	METEOR	ROUGE	METEOR	ROUGE	METEOR	ROUGE
TVL-LLaMA	0.0730	0.3116	0.0399	0.1890	0.0438	0.2044
GPT-4V	0.0246	0.0795	0.0523	0.2197	0.0491	0.2038
Clip-KNN-Vote	0.2706	0.3237	0.1385	0.1971	0.1511	0.2087
RAV(SyncVote)	0.3015	0.3447	0.1416	0.1979	0.1584	0.2143
RAV(DualVote)	0.2572	0.2935	0.1413	0.1951	0.1520	0.2076
RAV(WeightVote)	0.2419	0.3036	0.1335	0.1859	0.1477	0.2028

not only surpassed the Clip-KNN-Vote baseline and GPT4V but also significantly outperformed the task-specific trained model, TVL. The other two strategies, DualVote and WeightVote, also demonstrated competitive results, although their advantage was less pronounced than that of SyncVote, a difference we attribute to their heuristic parameter settings.

Dataset quality significantly influences model performance. SSVTP, with all tactile data containing valid contacts and fully manually annotated labels, offers the highest quality. HCT, with substantial shaking or non-contact data and GPT-4V pseudo-labeling, is of lower quality. TVL, a mixture of SSVTP and HCT, has intermediate quality. RAV's three voting strategies (SyncVote, DualVote, WeightVote) consistently achieve superior performance on the high-quality SSVTP dataset compared to HCT and TVL, aligning with trends observed in training-dependent models. This indicates that RAV's performance improves with higher dataset quality. Furthermore, incorporating a small amount of high-quality SSVTP data into the lower-quality HCT dataset significantly enhances RAV's performance without substantially increasing computational cost, offering insights for data efficiency optimization.

5 Conclusion

In this paper, we introduce RAV (Retrieval-Augmented Voting), a knowledge-based cross-modal perception model that distinctly differs from conventional models reliant on extensive trainable parameters. RAV achieves visual-tactile cross-modal perception through a parameter-free design. Specifically, during the knowledge construction phase, visual and tactile image features extracted by CLIP are stored in visual and tactile vector databases, $\mathcal{M}_{\mathcal{V}}$ and $\mathcal{M}_{\mathcal{T}}$, respectively. Upon receiving visual and tactile query inputs, relevant data are retrieved from $\mathcal{M}_{\mathcal{V}}$ and $\mathcal{M}_{\mathcal{T}}$, and fused via a voter to generate tactile descriptions.

The voting phase incorporates three strategies: SyncVote (balanced voting), DualVote (sensory credibility-based), and WeightVote (distance-weighted). To validate RAV's effectiveness, we compare it against large-scale models (e.g., GPT-4V, TVL-LLaMA) in the tactile description task, using the TVL dataset (comprising SSVTP and HCT subsets). Results demonstrate that RAV achieves performance comparable to the best generative models in accuracy, with particularly strong results on high-quality datasets (e.g., SSVTP). Its performance significantly improves with increased dataset quality and scale, owing to its parameter-free nature, which enables incremental data upgrades without additional training at minimal computational cost. This approach offers a new direction for efficient cross-modal perception.

Limitations

RAV excels in the tactile description task with low computational cost, yet its design and evaluation reveal limitations that guide future research. The performance depends on the quality of vector databases $\mathcal{M}_{\mathcal{V}}$ and $\mathcal{M}_{\mathcal{T}}$. The current experiments utilize the TVL dataset (SSVTP and HCT subsets), where SSVTP's high-quality manual annotations significantly enhance performance, whereas HCT's pseudo-labels and non-contact data reduce retrieval accuracy. If the vector databases contain more noise or fail to cover specific tactile scenarios (e.g., rare materials), RAV's cross-modal label fusion may be compromised. The CLIP encoder, is pre-trained on vision-language tasks and may not fully capture the nuanced semantics of tactile data. Although RAV mitigates some feature limitations through voting strategies, a tactile-optimized en-

coder could further enhance performance. Additionally, the voting strategies (SyncVote, DualVote, WeightVote) rely on sensory credibility and distance-based weights, their robustness in complex scenarios, such as HCT's non-contact data, is limited, particularly when retrieved labels have low credibility, potentially leading to fusion biases.

Funding

This work was supported by the Opening Project of Robotic Satellite Key Laboratory of Sichuan Province and 2035 Innovation Pilot Program of Sichuan University.

References

ajbarnett. 2024. 400 words to describe texture. https://owlcation.com/humanities/Describing-Texture-400-words-to-describe-texture.

Lasse Emil R. Bonner, Daniel Daugaard Buhl, Kristian Kristensen, and Nicolás Navarro-Guerrero. 2021. Au dataset for visuo-haptic object recognition for robots. *ArXiv*, abs/2112.13761.

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

Tim Dettmers and Luke Zettlemoyer. 2023. The case for 4-bit precision: k-bit inference scaling laws. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 7750–7774. PMLR.

Letian Fu, Gaurav Datta, Huang Huang, William Chung-Ho Panitch, Jaimyn Drake, Joseph Ortiz, Mustafa Mukadam, Mike Lambeta, Roberto Calandra, and Ken Goldberg. 2024. A touch, vision, and language dataset for multimodal alignment. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Ruohan Gao, Yiming Dou, Hao Li, Tanmay Agarwal, Jeannette Bohg, Yunzhu Li, Li Fei-Fei, and Jiajun Wu. 2023. The objectfolder benchmark: Multisensory learning with neural and real objects. *CoRR*, abs/2306.00956.

Yunhai Han, Rahul Batra, Nathan Boyd, Tuo Zhao, Yu She, Seth Hutchinson, and Ye Zhao. 2021. Learning generalizable vision-tactile robotic grasping strategy for deformable objects via transformer. *CoRR*, abs/2112.06374.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. 2023. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 13916–13932. PMLR.

Ziyan Jiang, Xueguang Ma, and Wenhu Chen. 2024. Longrag: Enhancing retrieval-augmented generation with long-context llms. *arXiv preprint arXiv:2406.15319*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Trans. Big Data*, 7(3):535–547.

Justin Kerr, Huang Huang, Albert Wilcox, Ryan Hoque, Jeffrey Ichnowski, Roberto Calandra, and Ken Goldberg. 2022. Self-supervised visuo-tactile pretraining to locate and follow garment features. *Robotics: Science and Systems XIX*.

Justin Kerr, Huang Huang, Albert Wilcox, Ryan Hoque, Jeffrey Ichnowski, Roberto Calandra, and Ken Goldberg. 2023. Self-supervised visuo-tactile pretraining to locate and follow garment features. In *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*.

Yuma Koizumi, Yasunori Ohishi, Daisuke Niizumi, Daiki Takeuchi, and Masahiro Yasuda. 2020. Audio captioning using pre-trained large-scale language model guided by audio-based similar caption retrieval. *Preprint*, arXiv:2012.07331.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. *Preprint*, arXiv:2303.16634.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Wei Li, Pan Zhang, Xiaoyi Dong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. 2024. RAR: retrieving and ranking augmented mllms for visual recognition. *CoRR*, abs/2403.13805.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9802–9822. Association for Computational Linguistics.

Nicolás Navarro-Guerrero, Sibel Toprak, Josip Josifovski, and Lorenzo Jamone. 2023. Visuo-haptic object perception for robots: an overview. *Auton. Robots*, 47(4):377–403.

Md. Rizwan Parvez, Wasi Uddin Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. 2021. Retrieval augmented code generation and summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2719–2734. Association for Computational Linguistics.

Haozhi Qi, Brent Yi, Sudharshan Suresh, Mike Lambeta, Yi Ma, Roberto Calandra, and Jitendra Malik. 2023. General in-hand object rotation with vision and touch. In *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*, volume 229 of *Proceedings of Machine Learning Research*, pages 2549–2564. PMLR.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2022. Retrieval-augmented transformer for image captioning. *Preprint*, arXiv:2207.13162.

Jinlin Wang. 2024. Vahagn: Visual haptic attention gate net for slip detection. *Frontiers in Neurorobotics*, 18.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2024. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. *Preprint*, arXiv:2211.06687.

Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. 2022. Touch and go: Learning from human-collected vision and touch. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Shu Zhong, Elia Gatti, Youngjun Cho, and Marianna Obrist. 2024. Exploring human-ai perception alignment in sensory experiences: Do llms understand textile hand? *CoRR*, abs/2406.06587.