VTMo: Unified Visuo-Tactile Transformer Encoder with Mixture-of-Modality-Experts

Zichen Zhang Peihao Li Yuan Cheng
University of Michigan
{zhangzzc, peihaoli, cherryc}@umich.edu

1. Introduction

Touch modality is one of the most essential ways humans interact with the physical world [5]. We engage with objects by *observing* and *touching* them. Developing a unified vision-touch model capable of processing both modalities can significantly advance autonomous agents, enabling interactions with the physical world like humans.

We propose the **V**isuo-**T**actile **Mo**del (VTMo), a modular Vision-Touch Transformer encoder designed to unify the strengths of dual-encoder and fusion-encoder architectures. By integrating the flexibility of dual-encoder models, which enable fast inference by pre-encoding features, and the accuracy of fusion-encoder models, which incorporate deep cross-modal interactions, as illustrated in Fig. 1, VTMo offers a robust solution for diverse cross-modal tasks. VTMo uses a shared self-attention mechanism combined with modality-specific and cross-modal experts. Each VTMo block routes inputs to three parallel expert networks for *vision*, *touch*, and *vision-touch*, facilitating modality-specific and cross-modal feature learning.

Due to its flexible architecture, VTMo can function as an image-only encoder, touch-only encoder, or vision-touch fusion encoder, making it versatile for tasks requiring either speed or accuracy. Testing the representations learned by VLMo on the Image-to-Touch Retrieval task, we show that our proposed method achieves comparative accuracy, is faster to train, and simultaneously requires less computation complexity. Implementation details are available at https://github.com/zichenzhang04/vtmo

2. Related Works

Dual-encoder. Recent advances in visuo-tactile modeling have explored approaches for aligning touch and vision embeddings. UniTouch [11] employs a *dual-encoder* architecture where touch and vision modalities are encoded separately, and cross-modal interaction is handled by ranking the cosine similarity between latent embeddings. While efficient at inference time due to pre-encoded features, dual-encoder architectures often underperform in tasks requiring deeper cross-modal understanding [4].

Fusion-encoder. An alternative approach is the *fusion-encoder*, which integrates touch and vision features through cross-modal attention, as seen in VilBERT [7] for vision and language. Fusion-encoder architectures are more effective in tasks requiring detailed cross-modal interactions but are computationally expensive because they necessitate jointly encoding all possible vision-touch pairs during inference.

Vision-Language Models (VLMs). VLMo [2] introduced a modular approach called mixture-of-modality-experts (MOME) to combine modality-specific and cross-modal features, inspiring the design of our method.

3. Method

VTMO Block. The architecture of VTMo follows the same design as BEiT-Base [1]. However, in each Transformer block [4, 10], following the MOME design [2], we replace the single feed-forward network in the standard Transformer block with a pool of three parallel modality experts, each of which is an independent feed-forward network, as shown in Fig. 2. These three experts each handle visual image encoding, tactile image encoding, and visual-tactile fusion. Given a previous block's output $\mathbf{H}_{l-1}$, the VTMo block calculates the output $\mathbf{H}_l$ by routing to a specific modality expert. Here, LN stands for layer normalization, and MSA is short for multi-head self-attention.

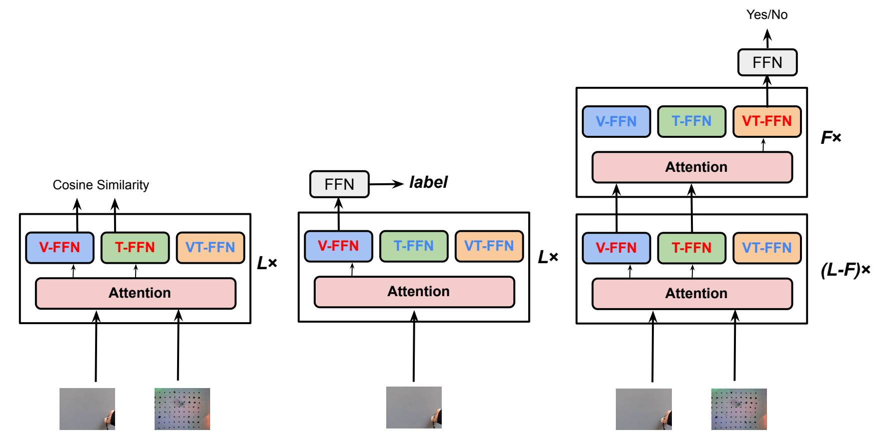
$$\mathbf{H}'_l = \mathrm{MSA}(\mathrm{LN}(\mathbf{H}_{l-1})) + \mathbf{H}_{l-1} \tag{1}$$

$$\mathbf{H}_l = \mathrm{Expert}(\mathrm{LN}(\mathbf{H}'_l)) + \mathbf{H}'_l \tag{2}$$

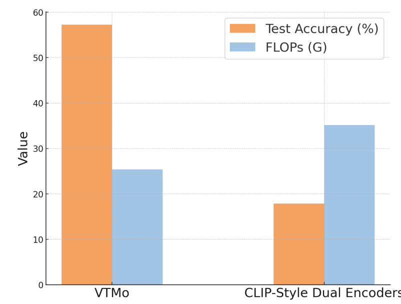
Input Representation. We treat tactile images the same as visual images. Following ViT [4], we obtain the standard patch embedding by linearly projecting both visual image patches and tactile image patches. We then employ the learnable special tokens `[I_CLS]` and position embeddings on both sequences of vision and touch.

Visual-Tactile Contrast. Inspired by contrastive learning methods [2, 8, 11], we design a visual-tactile contrastive

[1]This project was completed as the final project for EECS 442: Computer Vision at the University of Michigan.



(a) **Overview of the flexibility of VLMo**. Due to its modular structure, VTMo can be used as a **dual-encoder**, a **single-modality encoder**, and a **fusion-encoder**, respectively, depending on the downstream tasks without adjusting *any* parameter. Modality experts that are marked in blue are those that are not routed to during testing or inference.



(b) **Image-to-Touch Retrieval accuracy and FLOPs.** Used as a dual-encoder with shared attention layers, VTMo is more **accurate** and requires **less computation** than the baseline dual-encoder with separate attention layers.

Figure 1. VTMo can be adapted to different single-modal and multi-modal tasks while achieving comparative performance.

loss to align the representations of visual and tactile modalities. For each input pair, the `[I_CLS]` tokens are treated as representations for both the visual image and tactile image. The final contrastive loss is the average of image-to-touch and touch-to-image cross-entropy losses. See Appendix D for detailed mathematical definitions.

$$\mathcal{L}_{\text{contrastive}} = \frac{1}{2}(\mathcal{L}_{i2t} + \mathcal{L}_{t2i}). \tag{3}$$

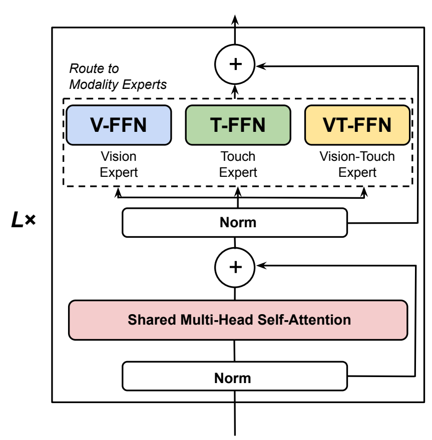


Figure 2. **Our proposed VLMo Transformer block.**

4. Image-to-Touch Retrieval

Training. Due to hardware limitations, we use a *randomly* sampled subset of the Touch and Go dataset [12] and a small batch size of 35. See Appendix B for more details. We structure VTMo as a dual-encoder following the left one in Fig. 1a. We initialize VTMo with the pre-trained weights from BEiT-Base-Patch16-224 [1]. The visual and tactile representations are aligned using the loss described in Appendix D, with a temperature parameter $\sigma = 0.07$. Since we noticed that freezing the attention layers decreases the performance (see Appendix C for ablation studies), we fine-tuned all parameters to ensure full adaptation to the new tactile modality. We use the Adam optimizer [6], with a learning rate of 1×10^{-4}. We train the model for 15 epochs. For baseline, we use the same setting to train two encoders similar to [8, 11], with the only difference being that the two encoders don't share attention layers.

Evaluation. We evaluate the model's performance on the test set using an image-to-touch retrieval task. Given an input visual image, the model retrieves the most closely aligned tactile image, as shown in Fig. 3. Retrieval accuracy is measured with Recall@1.

Results. As seen in Fig. 1b, our method achieves competitive performance and inference speed, while converging faster (see details in Fig. 4).

5. Conclusions and Future Work

In this work, we proposed VTMo, a unified visuo-tactile transformer encoder that leverages a mixture-of-modality-experts to balance efficiency and accuracy across single-modal and multi-modal tasks. While we demonstrated its effectiveness as a dual encoder for image-to-touch retrieval, future work includes evaluating VTMo as a fusion encoder and applying the learned representations to more challenging downstream tasks such as X-to-Touch generation and image synthesis with touch.

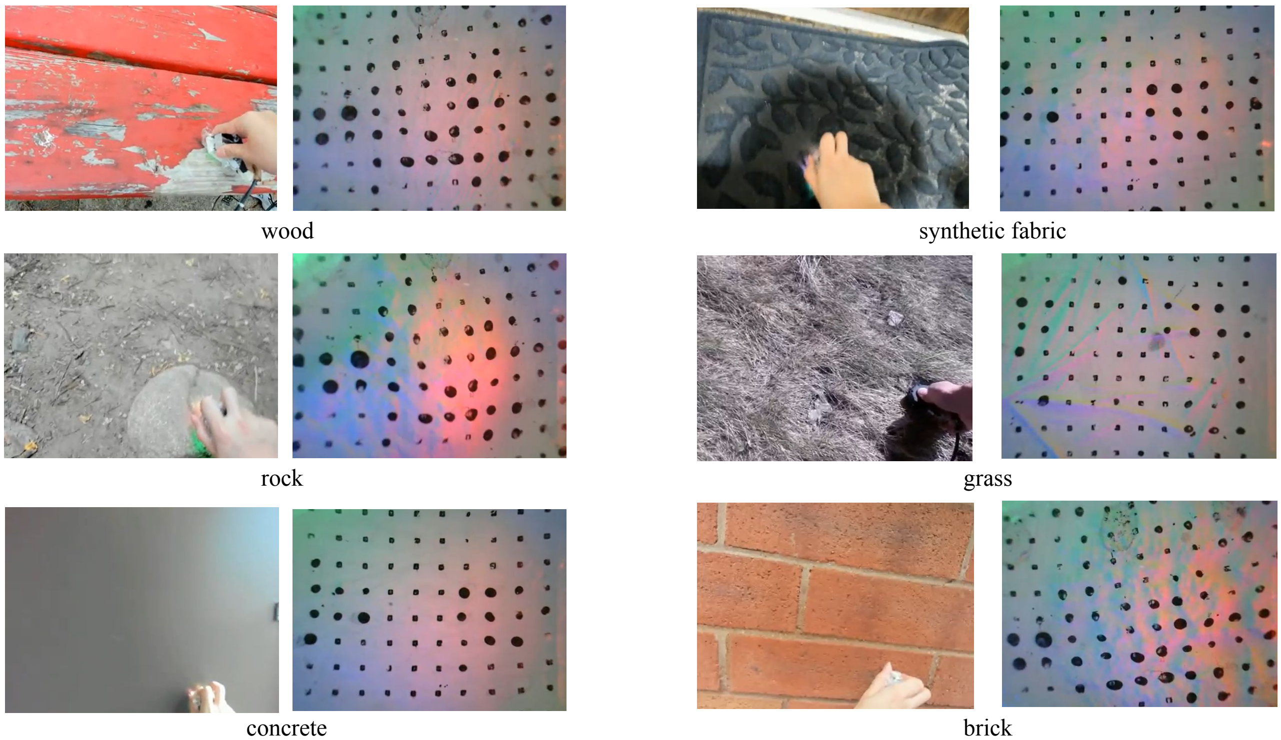


Figure 3. **Overview of Image-to-Touch Retrieval results.** Given an input visual image on the left, VTMo retrieves the most closely aligned tactile image, which is shown on the right.

References

[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022. 1, 2, 4

[2] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. VLMo: Unified vision-language pre-training with mixture-of-modality-experts. In *Advances in Neural Information Processing Systems*, 2022. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 1

[5] Fabian Hutmacher. Why is there so much more research on vision than on any other sensory modality? *Frontiers in Psychology*, 10, 2019. 1

[6] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 2

[7] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019. 1

[8] 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. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 1, 2

[9] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 4

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 1

[11] Fengyu Yang, Chao Feng, Ziyang Chen, Hyoungseob Park, Daniel Wang, Yiming Dou, Ziyao Zeng, Xien Chen, Rit Gangopadhyay, Andrew Owens, and Alex Wong. Binding touch to everything: Learning unified multimodal tactile representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26340–26353, June 2024. 1, 2

[12] Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learn-

ing from human-collected vision and touch. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 2, 4

A. Training

We found that VTMo converges much faster to a lower loss, as seen in Fig. 4.

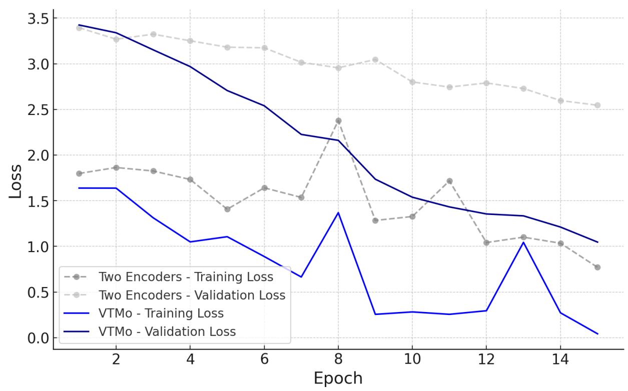


Figure 4. **Training loss in relation to the number of epochs.** VTMo converges faster than the baseline while achieving lower loss and better generalization.

B. Dataset Details

The subset of Touch and Go [12] we used includes a total of 3,620 pairs of visual and tactile images pairs. We split the dataset into three parts: 2,534 pairs for training, 543 pairs for validation, and 543 pairs for testing. Each visual-tactile pair represents a positive pair.

C. Ablation Studies

As shown in Tab. 1, we noticed that freezing the attention layers decreases the performance. We suspect that this performance gap results from the fact that tactile images were not represented in ImageNet-21k [3] that was used to pre-train BEiT [1].

Method	Accuracy
VTMo (frozen attention layers)	15.11
VTMo (no frozen layers)	**57.27**

Table 1. **Image-to-Touch Retrieval in relation to whether attention layers are frozen.** Fine-tuning all layers, including the shared attention layers (whose weights are initialized with BEiT-Base), achieves a much higher accuracy.

D. Contrastive InfoNCE Loss

Following [9], let $\hat{\mathbf{h}}_i^v \in \mathbb{R}^D$ and $\hat{\mathbf{h}}_j^t \in \mathbb{R}^D$ denote the normalized representations of the i-th visual image and j-th tactile image, respectively. The image-to-touch similarity $s_{i,j}^{i2t}$ and the touch-to-image similarity $s_{i,j}^{t2i}$ are calculated as:

$$s_{i,j}^{i2t} = (\hat{\mathbf{h}}_i^v)^\top \hat{\mathbf{h}}_j^t, \quad s_{i,j}^{t2i} = (\hat{\mathbf{h}}_i^t)^\top \hat{\mathbf{h}}_j^v. \tag{4}$$

To obtain the probability distributions for image-to-touch and touch-to-image matches, softmax normalization is applied over the respective similarities:

$$p_i^{i2t} = \frac{\exp(s_{i,i}^{i2t}/\sigma)}{\sum_{j=1}^N \exp(s_{i,j}^{i2t}/\sigma)} \tag{5}$$

$$p_i^{t2i} = \frac{\exp(s_{i,i}^{t2i}/\sigma)}{\sum_{j=1}^N \exp(s_{i,j}^{t2i}/\sigma)}, \tag{6}$$

where σ is a learnable temperature parameter shared across both modalities. The loss for aligning visual and tactile modalities is based on cross-entropy, calculated separately for image-to-touch and touch-to-image similarities:

$$\mathcal{L}_{i2t} = -\frac{1}{N}\sum_{i=1}^N \log p_i^{i2t} \tag{7}$$

$$\mathcal{L}_{t2i} = -\frac{1}{N}\sum_{i=1}^N \log p_i^{t2i}. \tag{8}$$

The total contrastive loss is then defined as Eq. (3).