Material Estimation using Audio-Visual cues on Multimodal LLMs

Dwait Bhatt
dhbhatt@ucsd.edu

Viraj Shah
vishah@ucsd.edu

Siddharth Satyam
ssatyam@ucsd.edu

Shashi Dhanasekar
sdhanasekar@ucsd.edu

Abstract

The ability to estimate the material composition of objects is crucial for humans, who rely on visual and auditory cues to inform how they handle objects with varying degrees of delicacy. Replicating this capability in robotic systems is essential for advancing their object manipulation skills. In this paper, we present a novel approach to material prediction that leverages both visual and audio data, integrating multiple modalities into a unified embedding space. Inspired by Macaw-LLM, our method comprises three key modules: the Modality Module, the Alignment Module, and the Cognitive Module. Using state-of-the-art models like CLIP and Whisper for encoding, we align multimodal features to a common textual embedding space, enabling both closed and open vocabulary material predictions. We demonstrate the effectiveness of our approach using the Greatest Hits dataset, showcasing improved accuracy and flexibility in material estimation for robotic applications.

1 Introduction

The ability to accurately estimate the material composition of objects is essential for humans, who often rely on visual and auditory cues to handle objects with appropriate care and precision. Replicating this capability in robotic systems is crucial for advancing their object manipulation skills, which are vital for a wide range of applications, from industrial automation to service robotics.

Material estimation in robotics not only impacts the system's ability to interact with its environment but also influences cost management, resource planning, project scheduling, and quality control. Accurate material estimation ensures financial feasibility, efficient resource allocation, and adherence to project timelines. Furthermore, it promotes sustainable practices by reducing waste and enabling the selection of environmentally friendly materials.

Traditional material estimation methods typically rely on single-modal inputs, such as visual or textual data, which may not capture the full spectrum of information necessary for precise predictions. To address these limitations, we propose a novel approach that leverages both visual and audio cues, integrating multiple modalities into a unified embedding space. This multi-modal approach aims to enhance the accuracy and flexibility of material prediction for robotic systems.

Our method is inspired by Macaw-LLM and consists of three major modules: the Modality Module, the Alignment Module, and the Cognitive Module. In the Modality Module, we utilize state-of-the-art models such as CLIP and Whisper to encode images and audio inputs, respectively. These encoded features are then aligned into a common textual embedding space through the Alignment Module, ensuring consistency across different modalities. Finally, the Cognitive Module employs a two-pronged strategy: a closed vocabulary model using a Multi-Layer Perceptron (MLP) for predefined

material classes and an open vocabulary model utilizing a Large Language Model (LLM) to generate more general text-based material predictions.

We validate our approach using the Greatest Hits dataset, which contains 977 videos of various objects being struck or scratched. This dataset provides a rich source of visual and auditory data, ideal for testing our multi-modal material prediction system. Our experimental results demonstrate that combining visual and audio cues significantly improves material prediction accuracy compared to using each modality independently. This work highlights the importance of integrating multiple modalities and advanced alignment techniques in enhancing the precision and flexibility of material estimation for robotics, ultimately contributing to more efficient and reliable robotic systems.

By leveraging the strengths of state-of-the-art encoding models and innovative alignment strategies, our approach represents a significant advancement in the field of robotic material estimation, paving the way for more capable and versatile robotic applications.

2 Related Works

2.1 Material Recognition

Initial attempts at material estimation involved using deep learning techniques to perform classification on the images of materials.

The authors of [10] used transfer learning for material classification. Transfer learning was used to analyze the contribution of shading information, reflectance and color to identify the main characteristics which determine into which material category an image belongs to. The information from the last convolutional layer provides information about the texture of the material which can be useful in determining the material of the object. The main conclusion one can draw from [10] is that object cues, such as shape and reflectance, are beneficial to material recognition, even probably essential.

In another work [1], the authors proposed a new texture descriptor, named FV-CNN, built on Fisher Vector (FV) pooling of a Convolutional Neural Network (CNN) local filter bank. FV pools local features densely within the described regions removing global spatial information, and is therefore more apt at describing materials than objects. FV is computed on the output of a single (last) convolutional layer of the CNN. The dense convolutional features are extracted at multiple scales and orderless pooled into a single FV. By doing so, the input image does not need to be rescaled to a specific size. This method incorporates multiscale information and describe regions of arbitrary shapes and sizes. The FV-CNN achieved 79.8% accuracy on FMD.

2.2 Multimodal models

Many papers have approached material recognition using classical as well as deep learning techniques. However, almost all of these papers are limited to using only images. With the huge increase in computational abilities, newer models like LanguageBind [12], LLaVA[4], and AudioCLIP[2] are able to integrate multiple modalities like audio, video and language. AudioCLIP extended CLIP to integrate audio, whereas LanguageBind and LLaVA open-sourced large multimodal models pre-trained on video-language pairs. Macaw-LLM [5] introduced an interesting concept of aligning the embeddings of multiple modalities into a common space.

3 Dataset

3.1 Greatest Hits

The dataset [6] contains 977 videos in which people hit or scratch different objects with a drumstick. This dataset was created to study the physical interactions within a visual scene.

The videos are captured from the viewpoint of an observer who is focused on the interaction taking place. This is particularly useful for our use case and aligns with our motivation. 64% of the videos are from an indoor setting and 36% from outdoor scenes. Each video, on average, contains 48 actions (approximately 69% hits and 31% scratches) and lasts 35 seconds.

The distribution of materials is presented in the form of a pie chart in the Figure 1

Each video has an associated text file that contains information about the timestamp of the action, the action and the material of the object being struck. We extract the image frames at each of these timestamps. We also extract a 1-second audio frame centered on the timestamp. We then convert the .wav file into mel-spectrogram to deal with storage limitations.

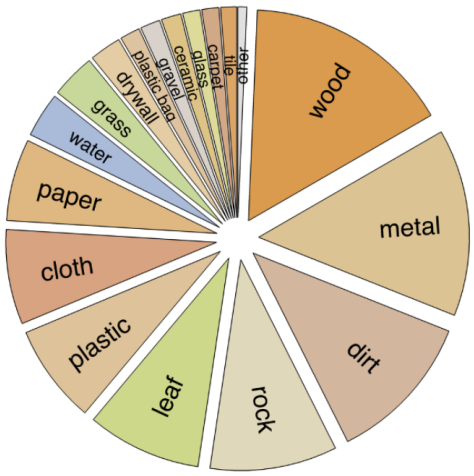


Figure 1: Pie chart depicting the distribution of materials in the Greatest Hits Dataset

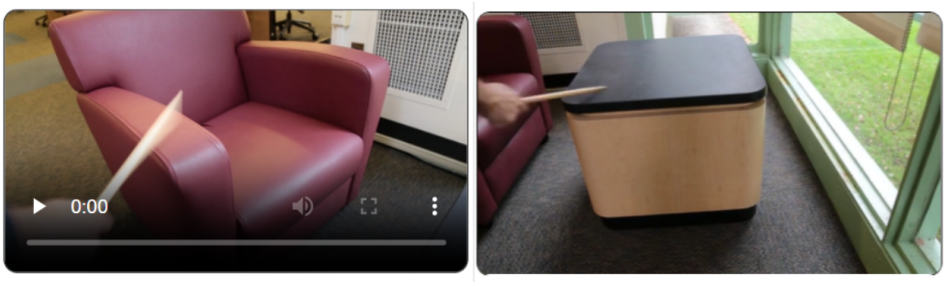


Figure 2: Greatest Hits Dataset

3.2 Flickr-Materials

Flickr-Materials dataset [9] was another dataset that was used. This dataset consisted of 100 images of 10 different materials. However, this dataset only had images and hence, only visual cues could be used for predicting the material.

Figure 3: Flickr Materials Dataset

4 Methodology

There are two baselines that are explored in this paper. First, using visual data only and second using audio data only. These baselines help in understanding the reason to use both visual and audio data.

4.1 Only Visual Cues

Initially, we evaluate how the models can use only visual cues to predict the material of an object. For this purpose, we make use of CLIP [7], Contrastive Language-Image Pre-training, which uses natural language to reference learned visual concepts resulting in zero-shot transfer of the model to downstream tasks like class prediction. A similarity score in the form of a dot product is computed between the image and its corresponding text to predict the label. The model is fine-tuned in a contrastive way using image-text pairs. Batches are created using a class-balanced sampler, which prevents having the same set of classes in a single batch. (with a batch size of 4)

4.2 Only Audio Cues

We evaluate the effectiveness of only audio cues in determining the material of an object being struck. We need to extract the audio features and pass them to a fully connected layer with softmax to classify the audio into one of the 17 classes of materials.

We extract the audio features using the VGGish [3] model, which is a pre-trained Convolutional Neural Network developed by Google. As the name suggests, the architecture of this network is inspired by the famous VGG networks used for image classification. It has been adapted and pre-trained on 60M AudioSet specifically for audio feature extraction. The network consists of a series of convolution and activation layers, optionally followed by a max pooling layer. This network contains 17 layers in total.

The input audio signal (.wav) is divided into 960-millisecond time blocks from which log mel-spectogram is extracted. This is fed into the model that consists of convolutional, activation and pooling layers. The output is a semantically semantically meaningful 128-D embedding vector.

4.3 Our Method

Inspired from Macaw-LLM [5], we define three major modules- i) Modality Module ii) Alignment Module, and iii) Cognitive Module. We used Image and Audio modalities to generate material predictions for closed and open vocabulary cases.

4.3.1 Modality Module

To handle multiple modalities effectively, we used CLIP [7] and Whisper [8] to encode Images and Audio inputs. We encode the multi-modal features as:

$$h_i = \text{CLIP}(x_i), \quad h_v = \text{CLIP}(x_v), \quad h_a = \text{WHISPER}(x_a), \tag{1}$$

where $h_i \in \mathbb{R}^{L_i \times d_h}$, $h_v \in \mathbb{R}^{L_v \times d_h}$ and $h_a \in \mathbb{R}^{L_a \times d_h}$ are image, video, and audio features, respectively, and d_h is the dimension of modality-specific features.

4.3.2 Alignment Module

Since the encoded vectors have unequal sequence lengths and embedding dimensions, they are first converted into the same dimensions. Despite being encoded through transformer-based encoders, the representations belong to different spaces and need alignment. This is done by aligning them to a common embedding space, which was chosen as the text-embedding space. The following equations give the alignment operation:

$$h_i' = \text{Linear}(\text{Conv1D}(h_i)), \quad h_v' = \text{Linear}(\text{Conv1D}(h_v)), \quad h_a' = \text{Linear}(\text{Conv1D}(h_a)), \tag{2}$$

where $h_i' \in \mathbb{R}^{L' \times d_e}$, $h_v' \in \mathbb{R}^{L' \times d_e}$, and $h_a' \in \mathbb{R}^{L' \times d_e}$ are the transformed features with a fixed length of L' and an embedding dimension of d_e.

To convert the encoded image and audio embeddings into the text embedding space, we refine the embeddings using cross-attention with the embedding matrix $E \in R^{V \times d_e}$ associated with a textual

LLM.

$$h^a = \text{Attn}(h', E, E), \tag{3}$$

where h' is the modality representation obtained in Equation 2 (i.e. h'_i, h'_v, and h'_a) and h^a is the corresponding aligned representation, specifically, h^a_i, h^a_v, and h^a_a.

4.3.3 Cognitive Module

Since the Greatest Hits data set contains 17 material classes, our first approach focuses on generating class predictions using an MLP in the cognitive module as shown in Fig.1.

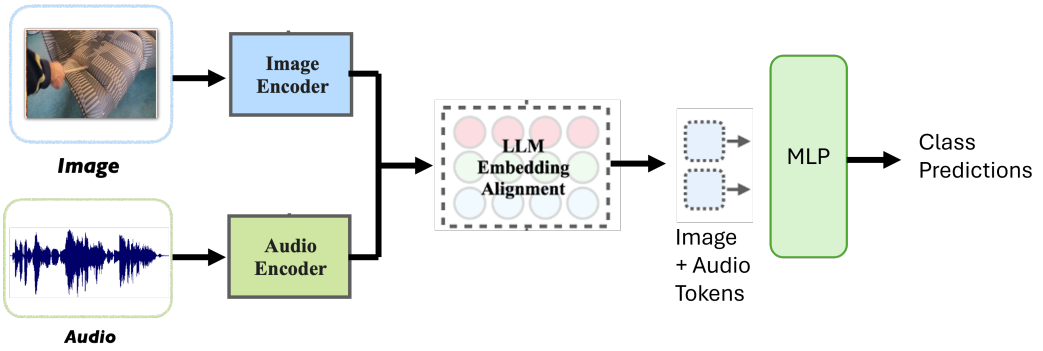


Figure 4: The closed vocabulary model to predict material classes.

We use a second approach, shown in Fig.2, which focuses on a more general text response using an LLM, which takes the aligned embeddings and an instruction as inputs and generates a textual response. The textual response helps in producing an open-vocabulary prediction for the material and is not restricted to the xx classes present in the dataset.

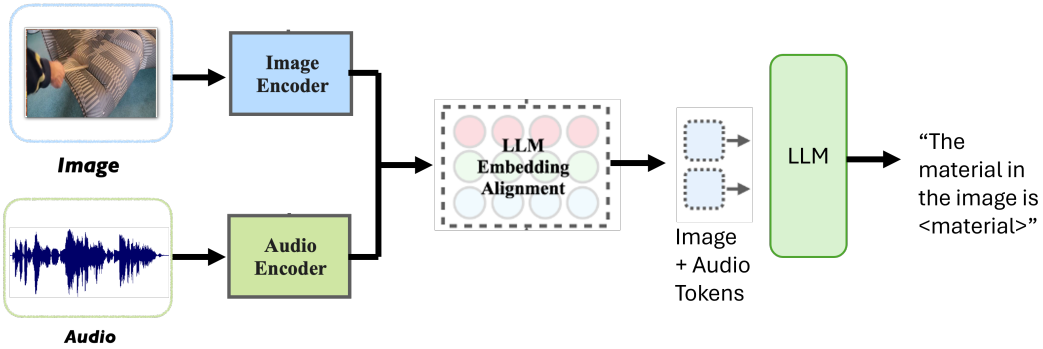


Figure 5: The open vocabulary model to predict materials as textual responses.

5 Experiments and Results

5.1 Only Visual Cues

We perform zero-shot prediction on Flickr Material Database (FMD) [9] that consists of the images of surfaces belonging to 10 materials. There are 1000 images in total. The model achieved an accuracy of 82% in the zero-shot inference.

Further, we fine-tuned the model on the FMD dataset and assessed its performance. The accuracy jumped to 85%.

We also perform zero-shot inference on the Greatest Hits Dataset with prompt engineering and evaluate the model.

The following are the prompts that we used:

(a) Plastic Frisbee that looks like ceramic plate

(b) Ceramic bowl that looks like glass

Figure 6: Deceptive Visual Cues

1. P0 - <material_name>
2. P1 - A picture of object made of <material_name>
3. P2 - A drumstick is hitting an object made of <material_name>

We can see an increasing trend in accuracy from P0 to P2. With prompt P2, the model achieved an accuracy of **35.39%**.

Again, we fine-tuned the model on the Greatest Hits Dataset and used the text with the best result from the zero-shot performance, which is P2. The accuracy jumped to **48.56%**. The results are presented in the Table 2.

The model achieves higher accuracy on the Flickr Materials Dataset since these images contain only the object of interest. Moreover, half the images in the dataset are close-up shots of different surfaces.

On the other hand, the Greatest Hits Dataset contains multiple objects, each being struck with a drumstick at different timestamps in the video. The frames extracted from the video also contain a brown drumstick. Most of the objects in the video have similar looks. In some instances, the objects have deceiving appearances that even humans might find hard to differentiate. This proves that additional information, like the audio associated with an object, might be a helpful feature. An example is shown in figure 6

Experiment	***Test Accuracy***
Zero-Shot	82%
Fine-tuned	**85%**

Table 1: Flickr Materials dataset

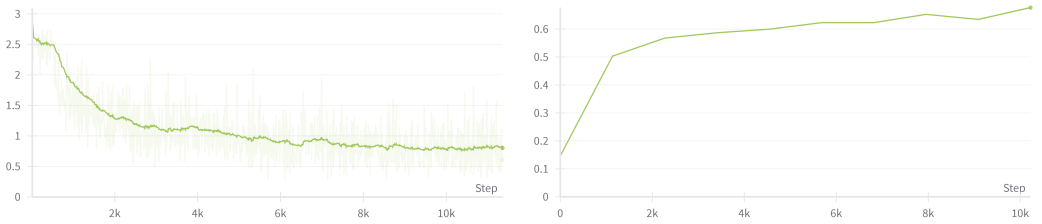
Experiment	***Test Accuracy***
Zero-Shot - P0	30.04%
Zero-Shot - P1	32.10%
Zero-Shot - P2	35.39%
Fine-tuned - P2	**48.56%**
Fine-tuned - P2 Train Accuracy	66.11%

Table 2: Greatest Hits dataset

Accuracy using Visual data only

5.2 Only Audio cues

We fine-tune the pre-trained VGGish + classification layer model on the Greatest Hits Dataset. The pre-trained network is kept static during training. Only the classification layer is updated. The performance was poor, with an accuracy of **10.7%** on the test data split of the Greatest Hits Dataset.



(a) Training loss

(b) Test accuracy

Figure 7: Training curves for multimodal model

The key insight that we can gain from the results of baseline models is that using visual and audio features independently is not useful in predicting the material of an object being struck. We can use them together to provide more meaningful context to the model for prediction.

5.3 Our Model

After training using our multimodal approach for closed vocabulary material estimation, we achieved a test accuracy of **67.69%**. This training run took over a day on a system equipped with 4 NVidia A40 GPUs. Figure 7 shows the training loss and accuracy plots. A comparison of the best results with different modalities is presented in Table 5.3. Our experiments clearly show that combining audio-visual information helps in material estimation. Due to resource and time constraints, we were unable to train our open vocabulary model. However, this remains a natural next step for our project.

Experiment	***Best Test Accuracy***
Audio-only	10.7%
Image-only	48.56%
Multimodal	**67.69%**

6 Conclusion and Future Work

State-of-the-art Multimodal models like Chat-GPT-4o do not perform well on the task of predicting the material of an object[1]. Models that are fine-tuned using audio and visual cues separately do not yield satisfying results. Our model combines audio and visual cues together and achieves an accuracy of 67.69% on fixed class material prediction tasks.

As a part of future work, we plan to perform open vocab material estimation by passing the audio and image encodings to an LLM with a suitable prompt. Additionally, to address the challenge of limited datasets containing images of objects, their materials, and audio of them being struck, we plan to explore training on synthetic audio data for known objects, such as those provided by [11].

References

[1] Mircea Cimpoi, Subhransu Maji, and Andrea Vedaldi. Deep filter banks for texture recognition and segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3828–3836, 2015.

[2] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022.

[3] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. Cnn architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017.

[1] Link to conversation

[4] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[5] Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration, 2023.

[6] Andrew Owens, Phillip Isola, Josh H. McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. Visually indicated sounds. *CoRR*, abs/1512.08512, 2015.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[8] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[9] Lavanya Sharan, Ruth Rosenholtz, and Edward H. Adelson. Accuracy and speed of material categorization in real-world images. *Journal of Vision*, 14(10), 2014.

[10] Patrick Wieschollek and Hendrik P. A. Lensch. Transfer learning for material classification using convolutional networks. *CoRR*, abs/1609.06188, 2016.

[11] Zhoutong Zhang, Jiajun Wu, Qiujia Li, Zhengjia Huang, James Traer, Josh H. McDermott, Joshua B. Tenenbaum, and William T. Freeman. Generative modeling of audible shapes for object perception. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[12] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023.