New York Smells: A Large Multimodal Dataset for Olfaction

Ege Ozguroglu[1] Junbang Liang[1] Ruoshi Liu[1] Mia Chiquier[1] Michael DeTienne[3]
Wesley Wei Qian[3] Alexandra Horowitz[1] Andrew Owens[2] Carl Vondrick[1]
[1]Columbia University [2]Cornell University [3]Osmo Labs
smell.cs.columbia.edu

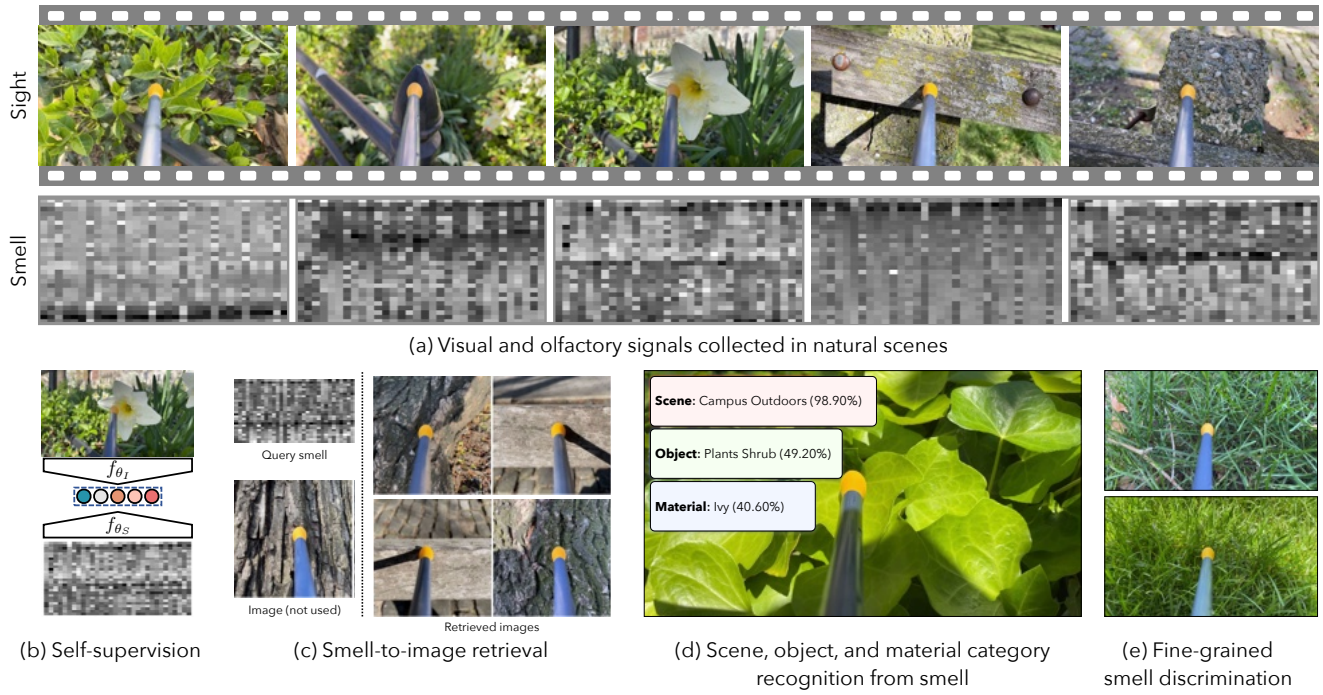


Figure 1. **Multimodal olfaction in-the-wild.** (a) We present *New York Smells*: a diverse, multimodal dataset of natural olfactory signals and paired visual data. We show one sequence of images and smell signals that we obtained in a public park (one scene of many in our dataset). We use this dataset for in-the-wild multimodal olfactory learning tasks that were not possible with previous datasets: (b) learning cross-modal features between olfaction and images, (c) retrieving images based on their corresponding olfactory signals, (d) recognizing in-the-wild scene, object, and material categories from smell, (e) distinguishing different grass species.

Abstract

While olfaction is central to how animals perceive the world, this rich chemical sensory modality remains largely inaccessible to machines. One key bottleneck is the lack of diverse, multimodal olfactory training data collected in natural settings. We present New York Smells*, a large dataset of paired image and olfactory signals captured "in the wild." Our dataset contains 7,000 smell-image pairs from 3,500 distinct objects across indoor and outdoor environments, with approximately 70× more objects than existing olfactory datasets. Our benchmark has three tasks: cross-modal smell-to-image retrieval, recognizing scenes, objects, and materials from smell alone, and fine-grained discrimination between grass species. Through experiments on our dataset, we find that visual data enables cross-modal olfactory representation learning, and that our learned olfactory representations outperform widely-used hand-crafted features.*

1. Introduction

Olfaction—the sense of smell—is a key way that animals, and to a lesser extent humans, perceive the world. Yet, this rich "chemical world", central to the sensory experience of many species, is largely imperceptible to machines.

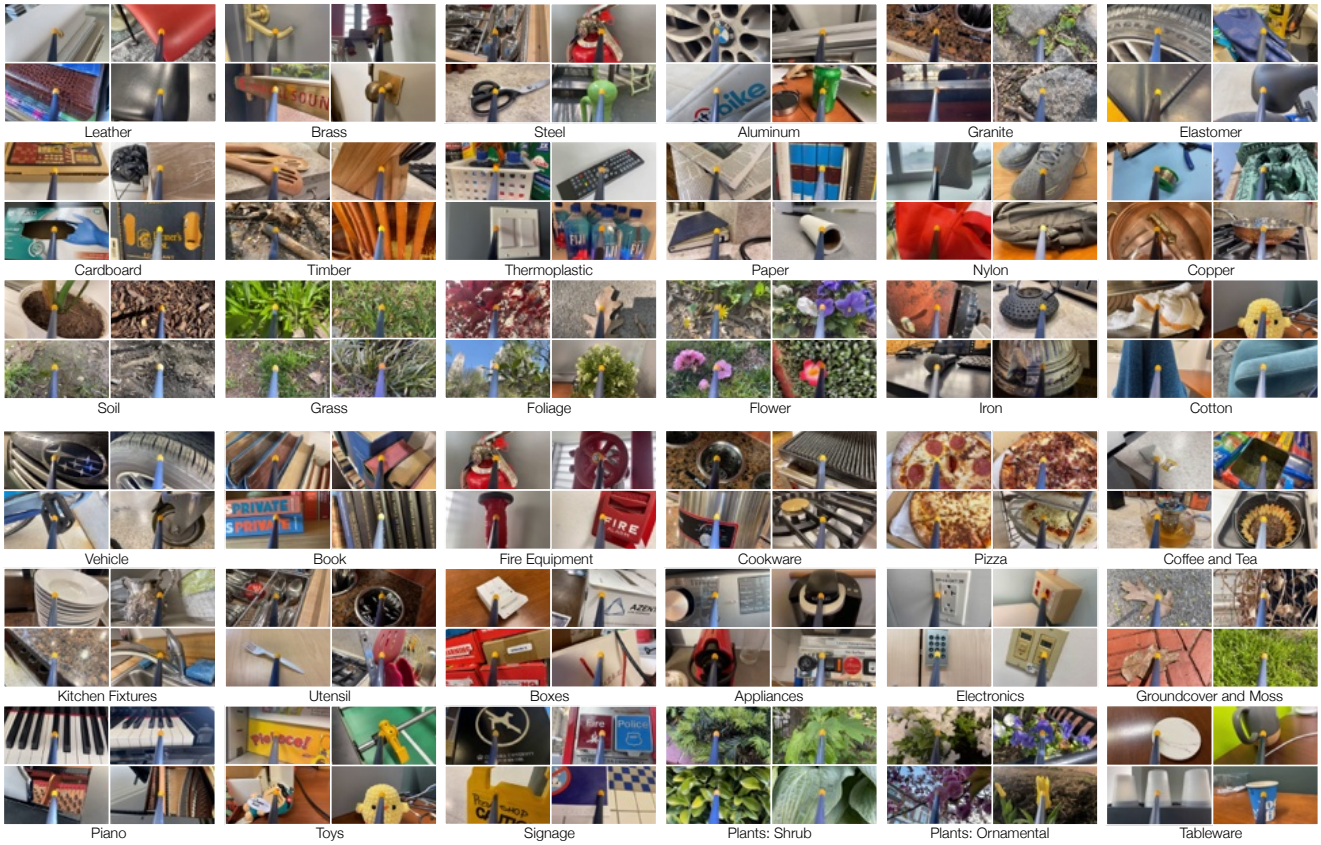


Figure 2. **The *New York Smells* dataset.** We collect a diverse dataset of paired sight and olfaction by visiting many locations within New York City and recorded a variety of materials (top rows) and objects (bottom rows) in different scenes. We show a selection of the captured images here. All samples have a corresponding olfactory signal captured from the Cyranose electronic nose.

This is in contrast to sight, sound, and touch, where advances in machine learning, particularly unsupervised and multimodal learning, have led to rapid improvements in machine abilities. One of the major obstacles to applying this approach to olfaction is the lack of suitable data. Existing olfaction datasets have largely been based on perceptual descriptors, rather than the raw outputs of olfactory sensors, or are captured in lab settings. Unlike audio and touch [8, 21, 46], existing olfactory datasets are not paired with vision (or other sensory modalities), making it difficult to link olfaction to the representations of other modalities.

In this paper, we address this problem by creating *New York Smells*, a large in-the-wild dataset of paired vision and olfaction. We visited dozens of indoor and outdoor scenes in New York City, such as parks, gyms, dining halls, libraries, and streets. We walked through each scene and recorded naturally synchronized images and smells of their odorant objects (Fig. 1a). To supplement this data, we also record a suite of other sensors: RGB-D, temperature, humidity, and volatile organic compound (VOC) concentration. Our dataset (Fig. 2), which contains 7000 olfactory-visual samples from 3500 objects, is significantly larger and more diverse than other olfactory datasets. For example, this is 70× as many distinct objects as the lab-collected concurrent work of Feng et al. [15].

We use our dataset's visual signals to establish a benchmark for in-the-wild smell perception. First, we propose a smell-to-image retrieval task that evaluates the ability of a model to establish cross-modal visual-olfactory associations (Fig. 1c). Second, we obtain pseudolabels for odorant objects, materials, and scenes, which we use to define corresponding category recognition tasks (Fig. 1d). Finally, we evaluate fine-grained recognition by proposing a benchmark for grass species recognition (Fig. 1e).

To further demonstrate the utility of paired visual-olfactory signals, we use our dataset for multimodal representation learning. Inspired by self-supervised learning methods in other multimodal domains [1, 32, 38], we train general-purpose olfactory features by training a joint embedding between smell and sight (Fig. 1b) using contrastive learning. Through experiments on our downstream tasks from our benchmark using a variety of different network architectures, we find that our learned olfactory representations significantly outperform hand-crafted smell features

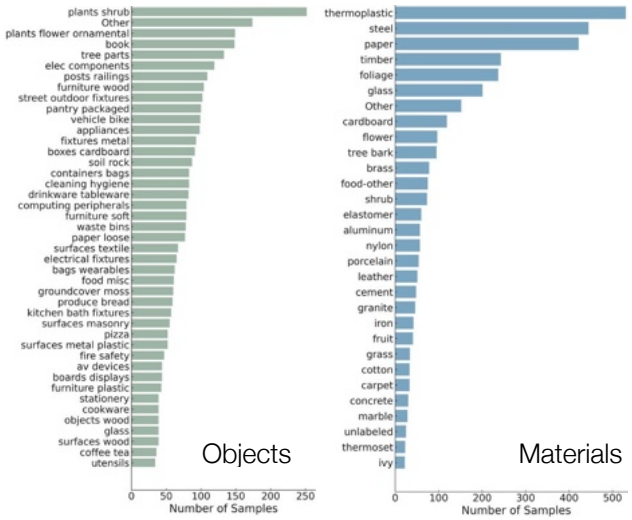


Figure 3. **Odorant analysis.** We show the distribution of objects and materials in our dataset. We use these labels to define smell understanding benchmarks.

that are widely used in prior work.

We see this dataset as a step toward in-the-wild, multimodal olfactory perception, as well as a step toward linking sight with smell. While olfaction has traditionally been approached in constrained settings, such as quality assurance, there are many applications in natural settings. For example, as humans, we constantly use our sense of smell to assess the quality of food, identify hazards, and detect unseen objects. Moreover, many animals, such as dogs, bears, and mice, show superhuman olfaction capabilities [25], suggesting that human smell perception is far from the limit of machine abilities.

Our work makes the following contributions:

- Our dataset provides much more diverse and naturalistic olfaction recordings than previous datasets.
- Our dataset is the first to pair in-the-wild olfaction with images.
- Using the visual signals in our dataset, we establish a benchmark for in-the-wild olfaction understanding.
- We show that vision provides supervision for general-purpose olfactory feature learning.

2. Related Work

Machine Olfaction. Previous work in machine olfaction have often focused on idealized settings, often requiring a miniature chemistry lab to be embedded into the system, which is expensive, bulky, and often impractical. Research has studied processing molecular structures for the purpose of designing scents [34] by using graph neural networks on synthetic data to predict human preferences such as perfume, or how to modify crops to prevent pests from spreading [18]. Other work does recognition from real-world sensors, e.g., detecting diseases like COVID [20] or explosive devices [39]. Despite these advances, machine olfaction research has been limited to a) narrow domains that lack the diversity and realistic complexity of everyday situations, and b) heavily relied on exact molecular information, which requires hardware that is not available to portable, low-cost sensors. Our dataset instead is designed for diverse scenarios, where sensors are noisy and incomplete, and methods must scale to vast domains of in-the-wild scents and odors.

Raw signals from electronic noses are high-dimensional and noisy, making data-driven methods attractive for uncovering structure. At the molecular level, psychophysical datasets enable models to predict perceptual attributes [24], and graph-based approaches propose a principal odor map (POM) [26]. Mixture studies are limited, showing approximate perceptual similarity [37] and the existence of olfactory metamers [33]. Exploratory work has used mass spectra [11], as well as ion-mobility and e-nose data [28], but largely under lab conditions. Recent work emphasizes the importance of calibrating olfactory neuroscience to natural concentration ranges [42], motivating the need for olfactory data in natural environments. In concurrent, unpublished work Feng et al. [15] collect a dataset of smells using an e-nose. However, their approach is limited to a highly controlled lab environment (they place one object at a time in a consistent room) and is relatively small scale (comprising 50 objects). By contrast, our work: 1) captures "in-the-wild" olfaction in natural environments of smells, 2) contains paired multimodal signals, 3) is much more extensive. We also go beyond prior work by using our dataset for multimodal representation learning.

We specifically focus on the Cyranose e-nose [35], since it is a popular, hand-held sensor that provides a rich olfactory signal that captures a variety of chemical properties. It has been applied to a range of scientific and industrial applications, such as measuring food quality [5, 27, 49], recognizing bacteria [4, 14], evaluating the quality of construction materials [2], detecting fires [30], monitoring wildlife and fauna [12, 17], and disease detection [36, 41].

Cross-Modal Supervision. There have been a variety of different methods for supervising one sensory modality using another. Early work by De Sa [10] proposed to use hearing to train vision through self-supervision. Ngiam et al. [29] used a deep generative model to learn an audio-visual speech representation. In contrast to these works, we use our dataset for *olfactory* representation learning through cross-modal supervision with sight. Our work is closely related to audio-visual [3, 31] and visual-tactile [13, 46, 48] data collection efforts in which a human probes objects with a sensor while recording video. By contrast, we pair olfaction with multiple visual sensors. Recent work has learned a multimodal representation of taste [6]. However, this

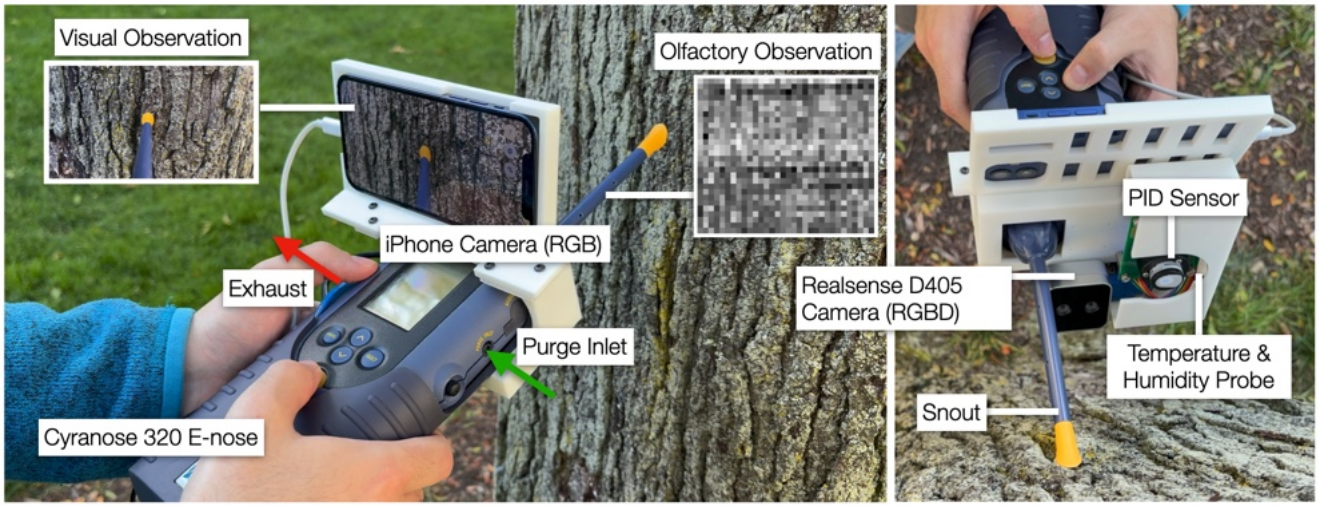


Figure 4. **Capturing paired sight and olfaction.** We walk through a variety of real-world scenes and capture paired olfaction and visual signals using a camera mounted to an e-nose on a custom 3D-printed sensor rig. We point the e-nose's snout at each object or substance of interest and record multiple images and smell signals from different orientations. We also capture a suite of other supplementary modalities: depth (from an RGB-D camera), temperature, humidity, and ambient VOC concentrations (from a PID sensor).

approach is based on solely on text descriptions of wine, whereas we use a real signal from a sensor.

Animal Olfaction. This work is motivated in part by the olfactory capacity of animals. Domestic dogs in particular, are renowned for having an extraordinary sense of smell. Their ability is manifest in various detection tasks, identifying everything from the presence of bed bugs to landmines to owners' low blood sugar [19]. Anatomically, dogs have hundreds of millions more olfactory receptor cells (the cells that begin the translation of VOCs into the perception of an odor) than humans do [22]. This enables them to detect more smells and more types of smells at lower concentrations. Their noses have separate routes for smelling and respiration, which enables airflow to arrive at the olfactory epithelium with every inhale [9]. The dog olfactory bulb is two percent of their brain by volume and sixty times the relative size of the human olfactory bulb [23].

3. The *New York Smells* Dataset

We collect a large-scale dataset of natural olfactory-visual sensory data. Specifically, our dataset contains multimodal "smell-centric" data. Unlike previous efforts on smell in machine perception [16, 26] and olfactory neuroscience [43], which rely on controlled or synthetic environments and stimuli, our dataset is collected in-the-wild. We probe everyday objects in their natural environments using paired vision and olfaction sensors. This approach captures the range of naturally occurring odorant concentrations, a property that is key for modeling olfaction under natural conditions [43]. We will publicly release the full dataset.

3.1. Collecting Natural Data

We now describe how we collected the dataset.

Hardware. To collect olfactory and visual data in natural environments, we leverage the natural synchronization between smell and sight during an olfactory observation. We chose to use the Cyranose 320 electronic nose [35], because it is a popular handheld sensor that is used in a wide variety of real-world smell sensing applications (see Sec. 2). Cyranose consists of a nanocomposite sensor array of 32 sensors. Each sensor responds to different chemical properties of volatile compounds that make up smell, without being specific to one volatile compound. We mount an iPhone 12 camera on Cyranose, angled to view the snout, where the olfactory measurement is collected. Cyranose operates at 2Hz, providing a 32-dimensional olfactory measurement at each timestep. Synchronized with the e-nose, a high-fidelity RGB camera captures the olfactory measurement at 1920×1080 resolution and 15 FPS.

We record an RGB-D signal using an Intel RealSense D405 (at 15 FPS and 1280×720 resolution), along with ambient temperature and humidity. We also collect complementary ambient volatile organic compound (VOC) concentrations using a MiniPID2 PPM WR sensor. The PID olfactory measurements reflect the naturally occurring concentrations of smells by diffusion, rather than active sniffing. The e-nose and all sensors are tethered real-time to a mobile station, consisting of battery, data storage, and compute to enable data collection across diverse settings, from parks, apartment settings, to streets. The complete capture set-up is shown in Figure 4.

The Cyranose obtains a measurement by actively draw-

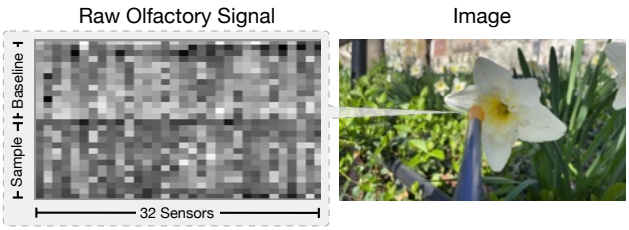


Figure 5. **Olfactory signal:** The raw smell signal is $T \times 32$ dimensions where T is the capture time. The first part of capture is the baseline phase, where the ambient background smell is sensed. The second part is the sample phase, where the smell of the object of interest is sensed. This example shows the response for a flower.

ing air through its snout and exposing the sampled compounds to an array of 32 sensors. Each sensor is a conductive polymer composite whose electrical resistance changes as odorant molecules are absorbed and cause the polymer to swell. We sample the resistance of all 32 sensors over time as the air sample is acquired, yielding a time-varying response for each channel. The raw olfactory signal is the matrix $x_S \in \mathbb{R}^{T \times 32}$, where each column corresponds to one sensor in the array and each row to one timestep.

Capturing procedure. Sensing objects in the scene requires separating the ambient odors from the odor of the object. For each sample, we first capture the baseline smell of the ambient environment, followed by the smell of the object of interest. Cyranose has two independent air pathways that can "sniff" outside air into its sensor chamber. The purge/baseline inlet, shown in Figure 4, on the side of Cyranose, pulls in ambient air, which leaves the sensor array through the exhaust outlet. Through this purge inlet-to-exhaust pathway, we first record the baseline smell for 10 seconds, receiving a 14×32 baseline matrix, representing each sensor for 14 timesteps. During this interval, air is drawn through the side port rather than the measurement inlet to avoid contamination from the target. Next, we record two samples through the main pathway, the snout. For data efficiency, we record two samples for each object from different positions. Both samples are 10 seconds. The raw olfactory data is thus a 28×32 matrix, which is the concatenation of the baseline and sample stages. See Figure 5.

Labeling the dataset. Figure 2 visualizes the scale and diversity of our vision & olfaction dataset. We used VLMs and the images in our dataset to automatically label the objects and materials. For materials, we used the Matador visual taxonomy of materials [7]. Using both views available in our dataset and this taxonomy as a closed set of categories, we generated material labels with VLMs (GPT-4o). For objects, we manually wrote a closed set of 49 categories that spans our dataset, then generating vision labels with VLMs (GPT-4o). We manually labeled the scene categories

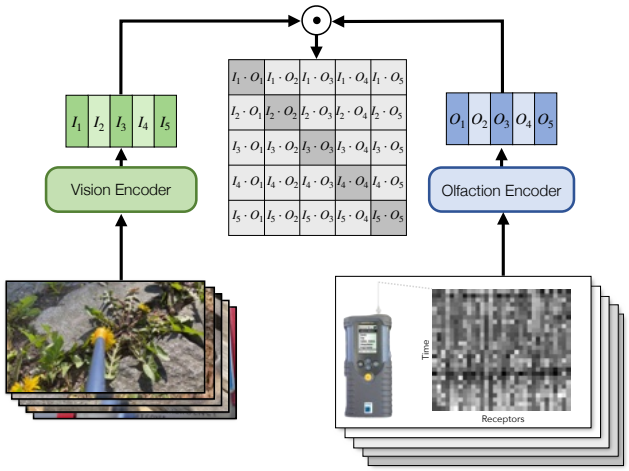


Figure 6. **Contrastive olfactory-image learning**. To demonstrate the effectiveness of our dataset, we train general-purpose olfactory representations using contrastive learning. We train the model to align co-occurring visual and smell signals. The visual encoder processes RGB images, while the olfaction encoder processes time-series sensor data from an e-nose.

for each sample, assigning each data collection session into one of 8 scene categories.

Dataset split. We uniformly split the dataset into train and validation splits. Since we collect two samples of each object during the capture procedure, we ensure that each sample appears in the same split, thus preventing overlap between the train and validation sets. The dataset has 7K olfactory-vision pairs, 3.5K unlabeled objects, 70 hours of raw video from both cameras, and 196K timesteps of raw smell measurement (baseline and sample stage olfactory measurements). Note that other dataset splits are possible as well, which we leave to future work, for example, studying generalization to novel ambient situations.

Dataset analysis. Figure 2 shows qualitative examples from the dataset and Figure 3 shows the distribution of materials and objects. We collected data across 60 sessions over two months. The dataset settings include several parks, university buildings, offices, streets, libraries, apartment settings, and dining halls. Each location had multiple data collection sessions. Our dataset has 41% outdoor and 59% indoor environments.

IRB. The Columbia IRB reviewed the collection procedure and determined it to be not human subjects research. No personal identifying information was collected.

4. Methods

As an application of our dataset, we use the multimodal synchronization between vision and olfaction to learn self-supervised representations for olfaction. The resulting representation can be used for cross-modal retrieval as well as

Figure 7. **Cross-modal retrieval qualitative results.** We use our joint embeddings to match smell to images. Given a query smell, we find the images in the dataset that are the closest match in embedding space. Each row shows a reference smell query along with the top 5 image retrievals predicted by our model. The ground-truth smell-image pair is highlighted in green.

Smell Encoder	**Mean Rank** ↓	**Median Rank** ↓	**Recall @ 5 (%)** ↑	**Recall @ 10 (%)** ↑	**Recall @ 20 (%)** ↑
Chance	467	467	0.536	1.07	2.14
MLP (Smellprint)	375.9	329	2.04	3.43	6.22
CNN (Raw Smell)	118.4	41	12.9	21.1	32.6
MLP (Raw Smell)	159.5	56	**17.3**	24.2	33.8
Transformer (Raw Smell)	**104.0**	**28**	16.5	**29.6**	**43.1**

Table 1. **Cross-modal retrieval quantitative results.** We evaluate model using standard retrieval metrics (N=933).

classification tasks.

4.1. Multimodal Contrastive Learning

Humans have a limited sense of smell and there are relatively few words to describe smells compared to other senses [44, 47] (although see [45] for possible exceptions). This gap makes it challenging to establish the label taxonomy and gather annotations on the scale required for effective machine learning. We instead will learn olfactory representations from unlabeled examples, leveraging cross-modal associations between smell and sight. Inspired by Contrastive Language-Image Pretraining (CLIP), we use contrastive learning to train a joint embedding between smell and images, which we analogously term Contrastive Olfaction-Image Pretraining (COIP).

Given the dataset of smell and corresponding visual data $\{\mathbf{x}_S^i, \mathbf{x}_I^i\}_{i=1}^N$, we learn olfactory and visual representations f_{θ_S} and f_{θ_I} by jointly training both encoders using a contrastive loss [40]:

$$\mathcal{L}_{I,S} = -\sum_{i=1}^{N} \log \frac{\exp\left(f_{\theta_I}(\mathbf{x}_I^i) \cdot f_{\theta_S}(\mathbf{x}_S^i)/\tau\right)}{\sum_{j=1}^{N} \exp\left(f_{\theta_I}(\mathbf{x}_I^i) \cdot f_{\theta_S}(\mathbf{x}_S^j)/\tau\right)}, \quad (1)$$

where $\tau = 0.07$ is the temperature. We analogously define the smell to image loss $\mathcal{L}_{S,I}$, where the denominator sums over the visual modality. We minimize both losses to learn the representations f_{θ_I} and f_{θ_S}:

$$\arg\min_{\theta} \; \mathcal{L}_{I,S} + \mathcal{L}_{S,I}. \quad (2)$$

By associating sight and smell (Figure 6), we learn a representation that can support the downstream interpretation of olfactory stimuli for multiple tasks. We apply these learned representations to a variety of tasks: smell-to-image retrieval (Fig. 1c), recognizing scene, material, and objects (Fig. 1d), and fine-grained discrimination between grass species (Fig. 1e).

4.2. Input Signals and Architectures

We experiment with two different input signals in our dataset: a raw representation that has no pre-processing, and a hand-crafted feature space that is widely used in machine olfaction research.

Raw Signal. Firstly, we directly use the raw signal from the sensor, which is an $T \times 32$ matrix representing the resistance of the 32 sensors inside the Cyranose over T timesteps. We directly input this matrix into the neural network, which then does contrastive learning, and allows end-

Figure 8. **Recognizing scenes, objects, and materials from smell.** We show the top 3 predictions from linear probing on the smell encoder. The predictions are from smell alone and the image is shown for visualization purposes only. Predictions are organized by color: red indicates scene classification, green indicates object classification and blue indicates material classification.

		Scenes			Materials			Objects		
Method	**Input**	Scratch	SSL	Rand	Scratch	SSL	Rand	Scratch	SSL	Rand
Chance		12.5	12.5	12.5	1.9	1.9	1.9	2.0	2.0	2.0
MLP	Smellprint	42.2	32.5	31.4	3.84	2.0	6.0	3.3	4.96	5.9
Transformer	Raw smell	91.0	90.4	72.7	2.33	14.0	7.1	13.8	18.4	12.3
CNN	Raw smell	99.5	95.0	74.5	11.9	12.3	9.4	17.9	19.8	8.7

Table 2. **Recognizing odorants.** We evaluate classification accuracy at recognizing scenes, objects, and materials from smell alone. For each approach, we compare end-to-end learning with scratch initialization (Scratch), self-supervised representations with linear probes (SSL), and linear probe with random weights (Rand).

to-end learning. We experiment with both convolutional neural networks (CNNs) and transformers as the backbone. By learning representations on this raw signal, there is the potential to discover highly powerful representations for olfaction that outperform hand-crafted features.

Smellprint. Secondly, we compare using a hand-crafted olfactory feature called a smellprint, which is widely used in representing smell from a Cyranose sensor [14, 17, 36, 41]. We use it as a baseline for the feature encoding. The smellprint produces a 32-dimensional vector from the raw smell matrix, and it summarizes the sensor response to odorants relative to the ambient environment. However, it discards many signals from the raw input, such as the 2nd-order statistics (e.g., correlations between different sensors).

The smellprint is computed by applying Savitzky–Golay filtering (window length w, polynomial order p) independently to each sensor time series in both the baseline and sample stages. Let $R_{i,j}$ denote the (filtered) resistance of sensor $i \in \{1, \ldots, 32\}$ at time index j. Let B be the baseline indices, and $S = S_1 \cup S_2$ the union of the two sample windows. Smellprint features compute per-sensor ambient level, sample peak, and the final feature as:

$$\text{Baseline (ambient) resistance: } R_{0,i} = \frac{1}{|B|} \sum_{j \in B} R_{i,j}, \quad (3)$$

$$\text{Sample peak resistance: } R_{\max,i} = \max_{j \in S} R_{i,j}, \quad (4)$$

$$\text{Smellprint (rel. response): } \mathrm{S}_i = \frac{R_{\max,i} - R_{0,i}}{R_{0,i}}. \quad (5)$$

We then directly feed the 32 dimensional vector in a multilayer percepton (MLP) before contrastive learning.

5. Tasks and Experimental Results

Our dataset evaluates three olfactory tasks: a) cross-modal retrieval between olfaction and vision, b) recognition tasks

including scene, material, and object classification from smell alone, and c) fine-grained discrimination of olfactory signals. We evaluate performance with supervised networks, contrastive unsupervised networks with linear probes, as well as the hand-crafted smellprint.

5.1. Cross-Modal Retrieval

Setup. For each query pair of smell and vision $\{\mathbf{x}_S^q, \mathbf{x}_I^q\}$ in our held-out test set, we sample a distractor set of images $D = \{\mathbf{x}_I^i\}_{i=1}^{N-1}$. We first embed the query pair into the shared olfactory and visual space to get $\{z_S^q, z_I^q\}$, where $z_S^q = f_{\theta_S}(\mathbf{x}_S^q)$ and $z_I^q = f_{\theta_I}(\mathbf{x}_I^q)$. We also embed every image $\mathbf{x}_I^i$ in D into the same olfactory-visual space: $z_i = f_{\theta_I}(\mathbf{x}_I^i)$. We sort every image feature z_i by its distance to the query smell feature z_S^q. If z_I^q is closest to z_S^q, then it will have a rank of 1. Following [32, 38], we use median rank, mean rank, and recall @ K to measure the percentage of smell queries for which the matching image embedding is ranked in the top K results.

Results. Table 1 compares CNN, MLP, and Transformer architectural variants of our olfactory encoder f_{θ_I} trained on the raw olfactory data as well as the hand-crafted smellprint. Contrastive pretraining using smellprint performs better than chance in all metrics. However, training the olfactory encoder on the raw olfactory signal leads to significant improvement compared to the smellprint encoder, independent of architecture. This shows the richer information present in the raw olfactory data, unlocking stronger cross-modal associations between sight and smell. We show qualitative results in Figure 7. Retrievals from our model often show semantic groupings. The odor of a book retrieves images of other books, the odor of leaves retrieves images of foliage. These results suggest that the learned representation captures meaningful cross-modal structure. Retrievals also group by material properties. For instance, the odor of moss on a concrete bench retrieves images of moss on tree bark and on another bench, while the odor of a wooden stick retrieves images of groundcover and a tree bark.

5.2. Scene, Object, and Material Recognition

Setup. We evaluate how well the representations are able to discriminate scenes, objects, and materials from smell alone. For each task, we compare architectural variants of the smell encoder (MLP, Transformer, CNN) trained via olfactory–visual contrastive learning against the same encoders with random weights, as well as versions trained on smellprint features rather than raw sensory inputs. We use linear probes on the olfaction representation. To train a probe, we use the activations from the penultimate layer of the olfaction network, and train it to predict the labels derived from the visual stream using GPT-4o on the training set. We then evaluate the probe on the held-out test set. Linear probes isolate the contribution of the representation itself. Strong linear probe performance indicates that the representation encodes semantic information that is linearly separable, and thus useful for downstream tasks.

Method	**Input**	**Accuracy**
Chance		50.0
Random weights	Smellprint	66.7
Trained from scratch	Smellprint	85.7
SSL + linear probe	Smellprint	90.0
Random weights	Raw smell	47.6
Trained from scratch	Raw smell	52.4
SSL + linear probe	Raw smell	**92.9**

Table 3. **Fine-grained discrimination**. We evaluate our olfaction models' ability to discriminate between grass species.

Results. As shown in Table 2, self-supervised olfaction representations trained with visual supervision outperforms baselines. Models trained on raw sensory inputs also achieve higher accuracy than models trained with the hand-crafted smellprint features. These results show that deep learning from raw olfaction signals is significantly better than hand-crafted features. In Figure 8, we showcase Top 3 predictions from linear probing our smell encoder, spanning diverse scenes, materials, and objects in our test set.

5.3. Fine-grained Discrimination

We ask whether learned olfactory representations can capture fine-grained differences, using our benchmark. In this benchmark, the goal is to distinguish between two grass species recorded at the same campus lawn, where they coexist. To test this, we collected alternating samples of both grass species across six 30-minute sessions, yielding a balanced dataset of 256 examples. We trained a linear classifier on the features learned through olfactory–visual contrastive learning and evaluated it on a held-out recording session of 42 samples.

Results. Table 3 shows classification accuracy for discriminating the two grass species. Training on the raw olfactory sensor signal (instead of hand-crafted features) yields the highest accuracy—exceeding all variants based on smellprints. These results suggest that olfactory–visual learning preserves more fine-grained information than learning with smellprints, and that visual supervision provides a signal for exploiting this information.

6. Conclusion

We present New York Smells, a real-world multimodal dataset of paired visual and olfactory signals collected in natural, in-the-wild environments. We demonstrated that visual data provides effective supervision for learning olfactory representations through contrastive learning, and mod-

els trained on raw olfactory signals substantially outperforming traditional hand-crafted features.

We see our work as opening two new research directions. It takes a step toward linking the fields of computer vision to computational olfaction, which have previously been studied separately. We have shown several ways that visual signals can supervise olfaction, such as through self-supervised contrastive learning with static images, but there are many other supervision cues that vision can provide, such as by conveying how objects change over time and 3D space. Our work is also a step toward creating olfactory methods that can successfully operate in-the-wild, rather than in lab settings. We will release code and data.

Acknowledgments

Funding for this research is provided in part by NSF Awards #2046910 and #2339071 and the NSF ERC for Smart Streetscapes. We thank Antonio Torralba for the early encouragement. We also thank Max and his large snout for much inspiration.

References

[1] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, 2017. 2

[2] Federico Autelitano, Erika Garilli, and Felice Giuliani. Electronic nose for smart identification of roofing and paving grade asphalt. *Transportation Research Procedia*, 40:4–11, 2019. 3

[3] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016. 3

[4] S Balasubramanian, S Panigrahi, CM Logue, M Marchello, and JS Sherwood. Identification of salmonella-inoculated beef using a portable electronic nose system. *Journal of Rapid Methods & Automation in Microbiology*, 13(2):71–95, 2005. 3

[5] Roberto Beghi, Susanna Buratti, Valentina Giovenzana, Simona Benedetti, and Riccardo Guidetti. Electronic nose and visible-near infrared spectroscopy in fruit and vegetable monitoring. *Reviews in Analytical Chemistry*, 36(4): 20160016, 2017. 3

[6] Thoranna Bender, Simon Sørensen, Alireza Kashani, Kristjan Eldjarn Hjorleifsson, Grethe Hyldig, Søren Hauberg, Serge Belongie, and Frederik Warburg. Learning to taste: A multimodal wine dataset. *Advances in Neural Information Processing Systems*, 36:7351–7360, 2023. 3

[7] M. Beveridge and S. K. Nayar. Hierarchical Material Recognition from Local Appearance. In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 5

[8] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 2

[9] Brent A. Craven, Eric G. Paterson, and Gary S. Settles. The fluid dynamics of canine olfaction: unique nasal airflow patterns as an explanation of macrosmia. *Journal of the Royal Society Interface*, 7(47):933–943, 2010. 4

[10] Virginia De Sa. Learning classification with unlabeled data. *Advances in neural information processing systems*, 6, 1993. 3

[11] Tanoy Debnath and Takamichi Nakamoto. Predicting human odor perception represented by continuous values from mass spectra of essential oils resembling chemical mixtures. *PLOS ONE*, 15(6):e0234688, 2020. 3

[12] Anna C Doty, A Dan Wilson, Lisa B Forse, and Thomas S Risch. Assessment of the portable c-320 electronic nose for discrimination of nine insectivorous bat species: implications for monitoring white-nose syndrome. *Biosensors*, 10 (2):12, 2020. 3

[13] Yiming Dou, Fengyu Yang, Yi Liu, Antonio Loquercio, and Andrew Owens. Tactile-augmented radiance fields. *arXiv preprint arXiv:2405.04534*, 2024. 3

[14] Ritaban Dutta, Evor L Hines, Julian W Gardner, and Pascal Boilot. Bacteria classification using cyranose 320 electronic nose. *Biomedical engineering online*, 1(1):4, 2002. 3, 7

[15] Dewei Feng, Carol Li, Wei Dai, and Paul Pu Liang. Smellnet: A large-scale dataset for real-world smell recognition. *arXiv preprint arXiv:2506.00239*, 2025. 2, 3

[16] Dewei Feng, Carol Li, Wei Dai, and Paul Pu Liang. Smellnet: A large-scale dataset for real-world smell recognition. *arXiv preprint arXiv:2506.00239*, 2025. 4

[17] Inês Ferreira, Teresa Dias, Juliana Melo, Abdul Mounem Mouazen, and Cristina Cruz. First steps in developing a fast, cheap, and reliable method to distinguish wild mushroom and truffle species. *Resources*, 12(12):139, 2023. 3, 7

[18] Ana Fundurulic, Jorge MS Faria, and Maria L Inácio. Advances in electronic nose sensors for plant disease and pest detection. *Engineering Proceedings*, 48(1):14, 2023. 3

[19] Simon Gadbois and Catherine Reeve. Canine olfaction: Scent, sign, and situation. In *Domestic Dog Cognition and Behavior: The Scientific Study of Canis familiaris*, pages 3–29. Springer, Berlin, Heidelberg, 2014. 4

[20] Christelle Ghazaly, Krystyna Biletska, Etienne A Thevenot, Philippe Devillier, Emmanuel Naline, Stanislas Grassin-Delyle, and Emmanuel Scorsone. Assessment of an e-nose performance for the detection of covid-19 specific biomarkers. *Journal of Breath Research*, 17(2):026006, 2023. 3

[21] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023. 2

[22] P. Hepper and D. Wells. Olfaction in the order carnivora: Family canidae. In *Handbook of Olfaction and Gustation*, pages 591–603. Wiley-Blackwell, Hoboken, NJ, 3 edition, 2015. 4

[23] P. G. Hepper. The discrimination of human odor by the dog. *Perception*, 17:549–554, 1988. 4

[24] Andreas Keller, Richard C. Gerkin, Yinzhao Guan, Avi Dhurandhar, Nath Tsaernstad, Jun Tan, Moustapha Bensafi, et al. Predicting human olfactory perception from chemical features of odor molecules. *Science*, 355(6327):820–826, 2017. 3

[25] Agata Kokcinska-Kusiak, Martyna Woszczylo, Mikolaj Zybala, Julia Maciocha, Katarzyna Barlowska, and Michal Dzieciola. Canine olfaction: physiology, behavior, and possibilities for practical applications. *Animals*, 11(8):2463, 2021. 3

[26] Brian K. Lee, Emily J. Mayhew, Benjamin Sanchez-Lengeling, Jennifer N. Wei, Wesley W. Qian, Kelsie A. Little, Matthew Andres, Britney B. Nguyen, Theresa Moloy, Jacob Yasonik, Jane K. Parker, Richard C. Gerkin, Joel D. Mainland, and Alexander B. Wiltschko. A principal odor map unifies diverse tasks in olfactory perception. *Science*, 381(6661):999–1006, 2023. 3, 4

[27] Changying Li, Paul Heinemann, and Richard Sherry. Neural network and bayesian network fusion models to fuse electronic nose and surface acoustic wave sensor data for apple defect detection. *Sensors and Actuators B: Chemical*, 125 (1):301–310, 2007. 3

[28] Philipp Müller, Katri Salminen, Ville Nieminen, Anton Kontunen, Markus Karjalainen, Poika Isokoski, Jussi Rantala, Mariaana Savia, Jari Väliaho, Pasi Kallio, Jukka Lekkala, and Veikko Surakka. Scent classification by k nearest neighbors using ion-mobility spectrometry measurements. *Expert Systems with Applications*, 115:593–606, 2019. 3

[29] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, Andrew Y Ng, et al. Multimodal deep learning. In *ICML*, pages 689–696, 2011. 3

[30] Ming Ni, Joseph R Stetter, and William J Buttner. Orthogonal gas sensor arrays with intelligent algorithms for early warning of electrical fires. *Sensors and Actuators B: Chemical*, 130(2):889–899, 2008. 3

[31] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H Adelson, and William T Freeman. Visually indicated sounds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2405–2413, 2016. 3

[32] 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, 2021. 2, 8

[33] Aharon Ravia, Kobi Snitz, Danielle Honigstein, Maya Finkel, Rotem Zirler, Ofer Perl, Lavi Secundo, Christophe Laudamiel, David Harel, and Noam Sobel. A measure of smell enables the creation of olfactory metamers. *Nature*, 588(7836):118–123, 2020. 3

[34] Benjamin Sanchez-Lengeling, Jennifer N Wei, Brian K Lee, Richard C Gerkin, Alán Aspuru-Guzik, and Alexander B Wiltschko. Machine learning for scent: Learning generalizable perceptual representations of small molecules. *arXiv preprint arXiv:1910.10685*, 2019. 3

[35] Sensigent. Cyranose 320 electronic nose, 2000. Product page. 3, 4

[36] W Christopher Shelley, Anthony R Pecoraro, Misty Good, Fikir M Mesfin, Krishna Manohar, John P Brokaw, Angela M Hansen, Robert H Pepin, Jonathan A Karty, Troy B Hawkins, et al. The impact of storage conditions on stool smellprints as assessed by an electronic nose. *ACS sensors*, 10(2):689–698, 2025. 3, 7

[37] Kobi Snitz, Ayelet Yablonka, Tal Weiss, Ilana Frumin, Rehan M. Khan, and Noam Sobel. Predicting odor perceptual similarity from odor structure. *PLOS Computational Biology*, 9(10):e1003184, 2013. 3

[38] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding, 2020. 2, 8

[39] Julio Torres-Tello, Ana Guaman, and Seok-Bum Ko. Mixed explosives dataset, 2019. 3

[40] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. 6

[41] Eva H Visser, Daan JC Berkhout, Jiwanjot Singh, Annemieke Vermeulen, Niloufar Ashtiani, Nanne K de Boer, Joanna AE van Wijk, Tim G de Meij, and Arend Bökenkamp. Smell—adding a new dimension to urinalysis. *Biosensors*, 10(5):48, 2020. 3, 7

[42] Matt Wachowiak, Adam Dewan, Thomas Bozza, T. O'Connell, and Elizabeth J. Hong. Recalibrating olfactory neuroscience to the range of naturally occurring odor concentrations. *Journal of Neuroscience*, 45(10):e1872242024, 2025. 3

[43] Matt Wachowiak, Adam Dewan, Thomas Bozza, Tom F O'Connell, and Elizabeth J Hong. Recalibrating olfactory neuroscience to the range of naturally occurring odor concentrations. *Journal of Neuroscience*, 45(10), 2025. 4

[44] Leo Weisgerber. Das geruchssinn in unseren sprachen ("the sense of smell in our languages"), 1928. 6

[45] Ewelina Wnuk, Rujiwan Laophairoj, and Asifa Majid. Smell terms are not rara: A semantic investigation of odor vocabulary in thai. *Linguistics*, 58(4):937–966, 2020. 6

[46] Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. *arXiv preprint arXiv:2211.12498*, 2022. 2, 3

[47] Yaara Yeshurun and Noam Sobel. An odor is not worth a thousand words: from multidimensional odors to unidimensional odor objects. *Annual review of psychology*, 61(1):219–241, 2010. 6

[48] Wenzhen Yuan, Shaoxiong Wang, Siyuan Dong, and Edward Adelson. Connecting look and feel: Associating the visual and tactile properties of physical materials. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5580–5588, 2017. 3

[49] Molin Zhou, Ragab Khir, Zhongli Pan, James F Campbell, Randall Mutters, and Zhuoyan Hu. Feasibility of detection of infested rice using an electronic nose. *Journal of Stored Products Research*, 92:101805, 2021. 3

7. Appendix

7.1. VLM Prompt for Labeling

The following Python function is used to label objects using GPT-4o, where images are passed to GPT-4o along with a structured prompt to select the closest matching object category.

Listing 1. Object labeling with GPT-4o.

```
def label_gpt_views(image_path1, image_path2, image_path3, image_path4, indexed_labels, labels):
    image_data1 = image_to_base64(image_path1)
    image_data2 = image_to_base64(image_path2)
    image_data3 = image_to_base64(image_path3)
    image_data4 = image_to_base64(image_path4)
    text_prompt = (
        "You are shown four images where a blue sensor probe with a yellow tip is pointing at the same object. "
        "from different angles. This is the same target object being analyzed.\n\n"
        "Choose the best matching category label for this object from the list below.\n\n"
        "Respond with the NUMBER corresponding to the best label. Do not invent labels. "
        "If none are perfect, choose the closest match.\n\n"
        "If the image is gray (with white plus sign), choose unlabeled (number 0).\n\n"
        "Category options:\n" +
        "\n".join(indexed_labels) + "\n\n"
        "Respond with only the number. No text."
    )
    response = client.chat.completions.create(
        model="gpt-4o",
        messages=[
            {
                "role": "user",
                "content": [
                    {"type": "text", "text": text_prompt},
                    {"type": "image_url", "image_url": {
                        "url": f"{image_data1}"
                    }},
                    {"type": "image_url", "image_url": {
                        "url": f"{image_data2}"
                    }},
                    {"type": "image_url", "image_url": {
                        "url": f"{image_data3}"
                    }},
                    {"type": "image_url", "image_url": {
                        "url": f"{image_data4}"
                    }},
                ]
            }
        ],
        max_tokens=10,
        temperature=0
    )
    label_number = int(response.choices[0].message.content.strip())
    return label_number
```

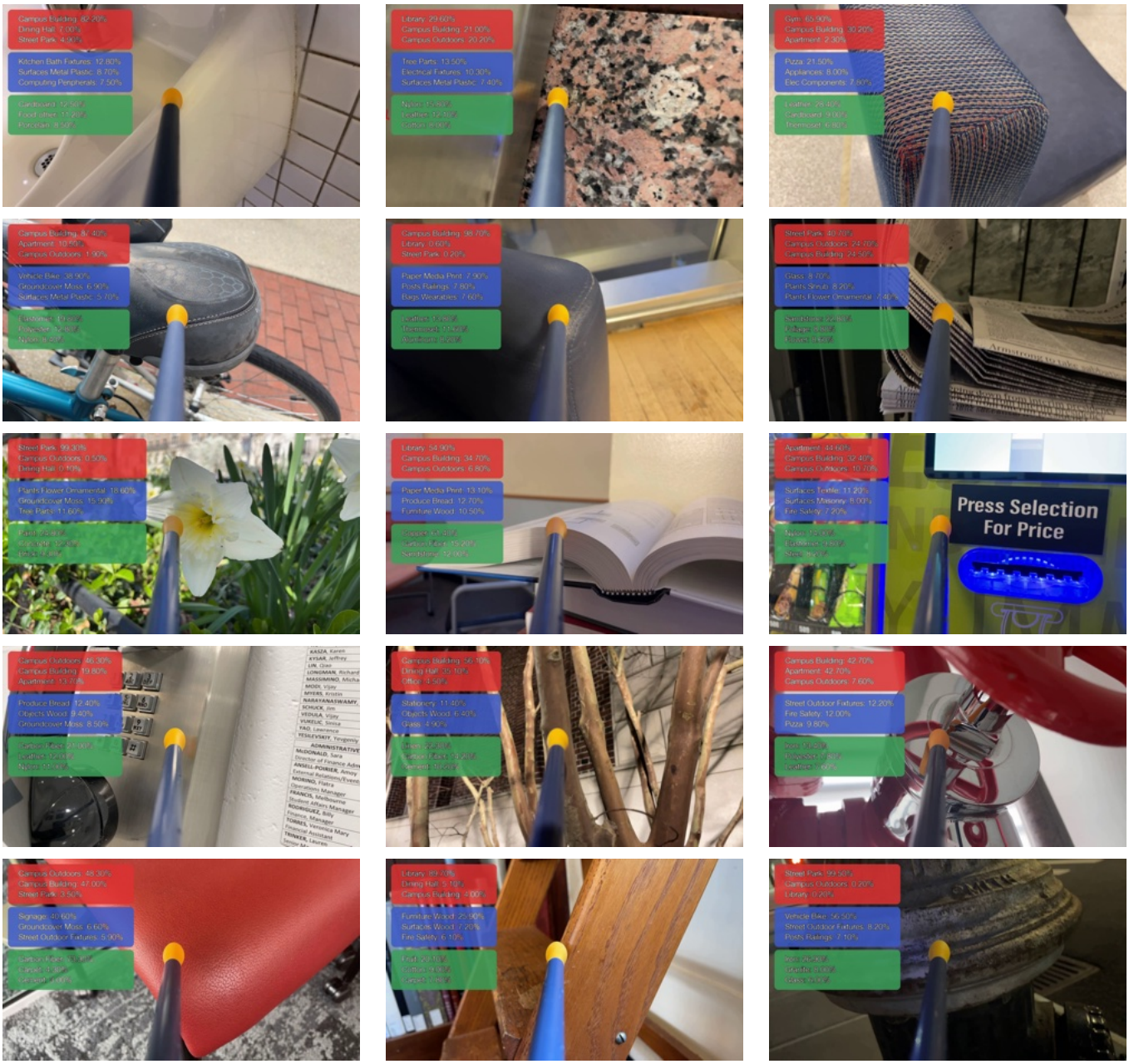
The following Python function sends two image views of the same object to GPT-4o to determine the object's underlying physical material. The prompt includes detailed instructions and examples to avoid visual or semantic biases.

Listing 2. Material labeling from two image views using GPT-4o.

```
def label_gpt_views(image_path1, image_path2, indexed_labels):
    image_data1 = image_to_base64(image_path1)
    image_data2 = image_to_base64(image_path2)
    text_prompt = (
        "You are shown two images where a blue sensor probe with a yellow tip is pointing at the same object. "
        "from different angles. This is the same target object being analyzed.\n\n"
        "Your task is to identify the object's **main physical material** - not what it contains, what it's shaped like, or what it's used for.\n\n"
        "Choose the most appropriate material from the following list:\n"
        f"{', '.join(indexed_labels)}\n\n"
        "**Label what the whole object is actually made of.** Ignore gloss, color, texture, logos, or symbolic cues.\n\n"
        "**Examples:**\n"
        "- Any food item other than bread -> 'food-other'\n"
        "- A smooth paper cup -> 'paper'\n"
        "- A shiny white bowl -> 'porcelain' or 'thermoplastic' (depends on shape, stiffness, and context)\n"
        "- A juice dispenser labeled 'orange juice' -> 'thermoplastic', not 'fruit'\n"
        "- A brown padded chair seat -> 'leather', not 'terracotta'\n"
        "- A light-colored sidewalk slab -> 'concrete' or 'cement', not 'asphalt'\n\n"
        "**Do not label based on:**\n"
        "- Color (e.g., orange != terracotta; gray != asphalt)\n"
        "- Function (e.g., juice bottle != fruit)\n"
        "- Gloss (e.g., shiny surface != glass)\n"
        "- Shape (e.g., cup shape != plastic)\n"
        "- Logos or printed text\n"
        "- Any object not being directly probed\n\n"
        "If you are absolutely certain it doesn't belong to any of the materials in the list, choose unlabeled. No text."
        "Now return **only the index** of the best matching material from the list as a number, without quotes. Never return any text such as 'I don't know' or 'unlabeled'."
    )
```

7.2. Additional Examples of Recognition from Smell

We include additional sampled examples showcasing the ability of our smell encoder to recognize scenes, objects, and materials from olfactory input alone. These examples further demonstrate the generalization of our model across diverse contexts in the test set and highlight the semantic structure captured by olfactory representations trained with visual supervision. Each prediction is obtained via linear probing and is color-coded by task type: red for scenes, blue for objects, and green for materials.



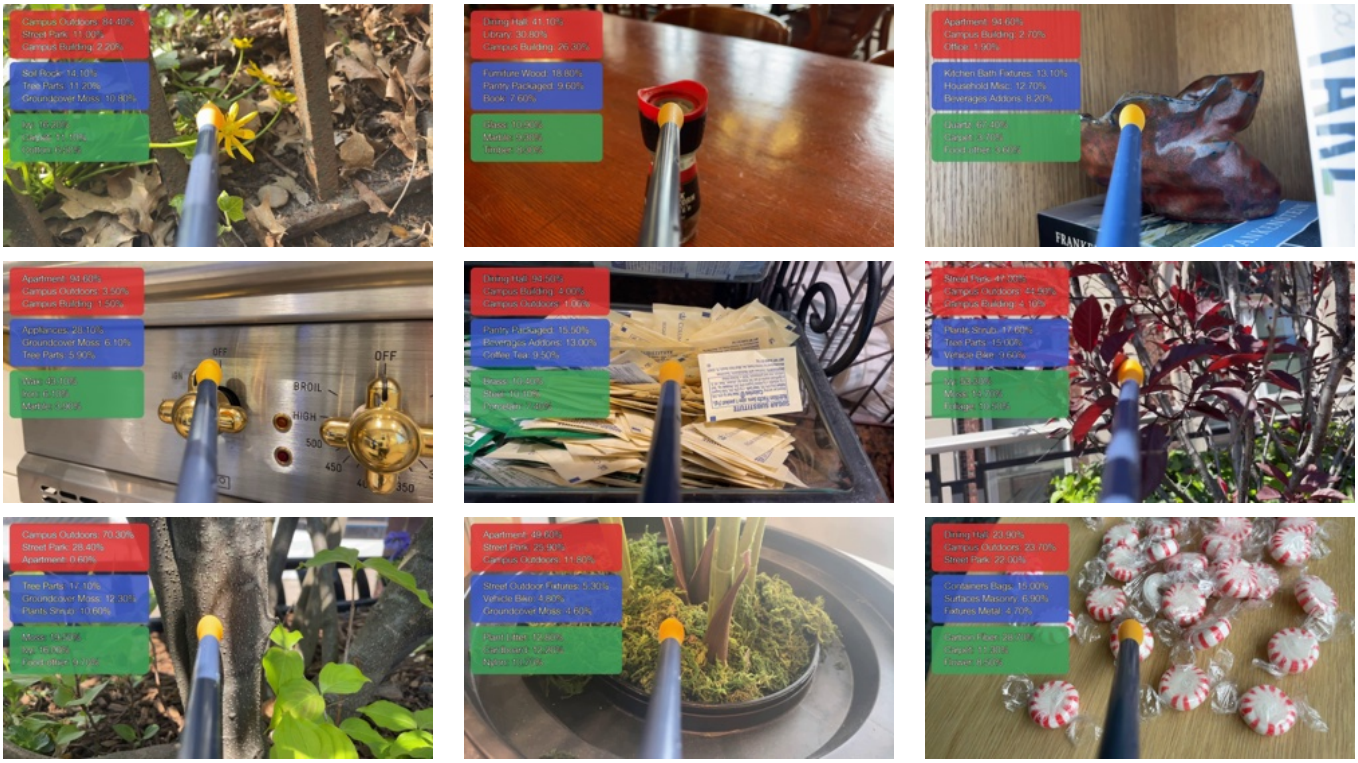


Table 4. Non-cherry-picked examples of recognizing scenes, objects, and materials from smell. Each tab displays the top-3 predictions obtained via linear probing on the smell encoder. Prediction types are color-coded: red for scenes, blue for objects, and green for materials.