

See & Sniff: Learning Visuo-Olfactory Representations

Seongyu Kim^{*1}, Seungwoo Lee^{*1}, Hyeonggon Ryu², Joon Son Chung¹, and Arda Senocak³

¹ Korea Advanced Institute of Science and Technology, Korea

² Hankuk University of Foreign Studies, Korea

³ Ulsan National Institute of Science and Technology, Korea

<https://mm.kaist.ac.kr/projects/SeeandSniff>

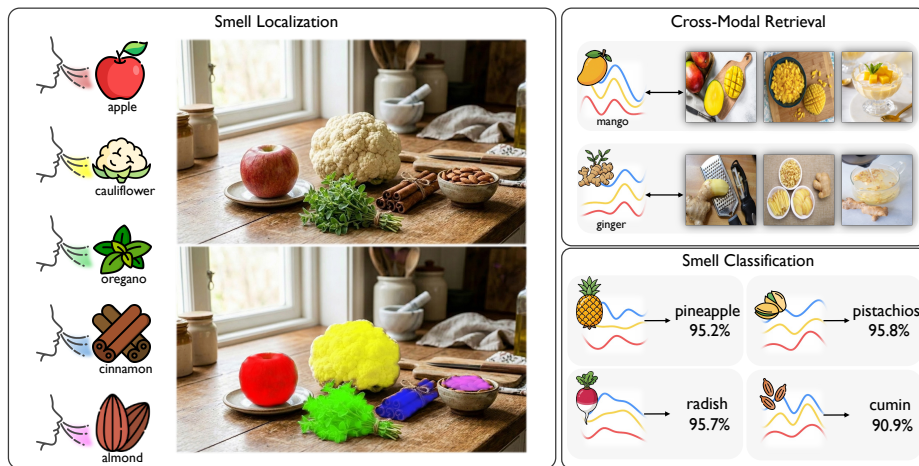


Fig. 1: What can *See & Sniff* do? We show that *See & Sniff*, the framework that learns joint visuo-olfactory representations, can handle both unimodal and cross-modal tasks. (Left) Smell localization identifies the locations of smell sources within a visual scene based on input olfactory signals. (Right Top) Cross-modal retrieval demonstrates semantic alignment through bidirectional retrieval. (Right Bottom) Smell classification predicts ingredients from olfactory inputs.

Abstract. While modern multimodal models integrate vision with language, audio, or touch, olfaction remains largely unexplored due to the lack of paired visuo-olfactory data. We introduce *SmellNet-V*, a scalable visuo-olfactory dataset built on the insight that odor identity is largely invariant to visual transformations within a semantic category. This allows us to synthetically pair smell-only samples with semantically aligned in-the-wild web images, converting a unimodal olfactory dataset into a cross-modal benchmark without costly co-collection. Building on this dataset, we propose *See & Sniff*, a self-supervised framework that learns joint visuo-olfactory representations via dense local alignment and naturally produces smell saliency maps for spatial grounding of odor sources. We further introduce pixel-level smell localization task and a benchmark for evaluation. Our method surpasses smell-only baselines by

* Equal contribution.

7% in smell classification from smell alone and generalizes to cross-modal retrieval and smell localization, establishing visuo-olfactory learning as a new direction in multimodal perception.

1 Introduction

*Everything has its own odor, which is its soul—tree,
flower, soil, rain, burning wood.* *Helen Keller*

Modern multimodal AI systems integrate vision with language [20, 26, 27, 34, 51], audio [2, 3, 14, 15, 17, 32, 38], and even touch [13, 23, 29, 48–50], yet overlook another fundamental sensory modality: *smell*. Despite its central role in biological perception, olfaction remains minimally explored in AI research, “*a neglected treasure*” [22]. This gap comes primarily from practical constraints: unlike cameras or microphones, portable olfactory sensors have historically been scarce and confined to laboratory settings, and large-scale public datasets for data-driven learning have been lacking. Recent advances in portable chemical sensing now enable more accessible machine-readable olfactory signals, and the first large-scale smell dataset has very recently been released [11]. However, it contains only olfactory measurements without paired visual data.

Why does visuo-olfactory integration matter? Human vision and olfaction are tightly intertwined sensory systems. When a person sees roasted coffee beans, they anticipate the characteristic bitter and roasted aroma even without smelling it. Conversely, the scent of coffee alone can evoke mental imagery of dark roasted beans or a steaming cup. These correspondences reveal a tight coupling between visual appearance and olfactory identity. Modeling this visuo-olfactory alignment computationally is central to our work.

Our objective in this work is to enable a self-supervised model to learn joint representations between visual and olfactory modalities, which fundamentally requires paired visuo-olfactory data. However, such datasets do not currently exist. We address this gap through *a key insight*: olfactory identity is largely invariant to visual transformations, including changes in lighting, scale, or minor color variations. For example, intra-class variations such as *a small red apple* and *a large red apple* emit nearly identical volatile organic compounds (VOCs) and are perceptually categorized by humans under the same odor category, ‘*apple*’. From a representation learning perspective, this suggests that visually diverse instances within a category should share a common semantic odor embedding. Leveraging this invariance, we construct a visuo-olfactory training dataset, *SmellNet-V*, by synthetically pairing each odor instance from an existing smell-only dataset [11] with semantically corresponding, in-the-wild open-world web images. This principled pairing strategy transforms a unimodal olfactory dataset into a scalable visuo-olfactory benchmark, enabling large-scale cross-modal alignment without the need for costly real-world paired data collection.

Multisensory perception extends beyond simple correspondence. Humans and other animals not only associate odors with visual identities, but also use olfactory cues to navigate and localize odor sources within complex environments [6, 35], *i.e.*, *smell localization*. To model this capability, we propose a

framework that computes dense similarity maps between local visual and olfactory features, enabling fine-grained alignment. The resulting smell saliency maps highlight image regions expected to emit a given odor, while simultaneously learning holistic cross-modal correspondences. To evaluate smell localization, we introduce the first dataset with pixel-level segmentation aligned with the odor samples.

Overall, our proposed training data, *SmellNet-V*, and self-supervised framework, *See & Sniff*, establish a new multimodal model for visuo-olfactory perception. The learned joint representations generalize across downstream tasks, including smell classification, cross-modal retrieval, and smell localization, as shown in Figure 1. In particular, for smell classification, our model surpasses the smell-only baseline by a significant margin, demonstrating that visuo-olfactory training enhances olfactory representation learning. Together, these contributions introduce new tasks and benchmarks for this emerging modality.

Our contributions are summarized as follows:

- We construct a visuo-olfactory dataset, *SmellNet-V*, by leveraging odor invariance within semantic categories to synthetically pair existing smell-only data with in-the-wild visual images.
- We propose *See & Sniff*, a self-supervised multimodal framework that learns joint visuo-olfactory representations via dense local alignment.
- We demonstrate that visuo-olfactory training enhances olfactory representation learning, surpassing the smell-only baseline by 7% in smell classification.
- We extend visuo-olfactory learning beyond classification and retrieval to spatial grounding by generating smell saliency maps and introducing the first pixel-level smell localization dataset for evaluation.

2 Related Work

Machine learning approaches to olfaction have primarily focused on predicting odor perception from chemical and molecular structure. Neural network-based models such as DeepNose [44] demonstrated that artificial networks can learn structured embeddings of odorant space. Subsequent works extended this direction using graph neural networks [1, 25, 37], multitask learning frameworks that capture shared representations across related odor categories [19], and attention-based aggregation mechanisms built upon molecular representations derived from chemical foundation models [21]. Other studies have addressed specific challenges such as odor intensity prediction [12]. Parallel efforts have explored artificial olfactory systems using sensor measurements. Electronic nose based frameworks, such as k-nearest neighbor-based scent classification method [31] and the eigengraph-based system [43], focus on signal-level representation learning and odor classification. While effective, these approaches remain unimodal and are primarily developed in controlled laboratory environments, limiting their scalability to large-scale real-world settings.

More recently, SmellNet [11] introduced the first large-scale real-world smell dataset collected with portable chemical sensors, along with ScentFormer, a

transformer-based model for olfactory representation learning. While this represents a significant step toward scalable data-driven machine olfaction beyond laboratory environments, the framework remains unimodal and lacks paired visual information. Building upon SmellNet, we extend its smell-only data to the visual domain by synthetically pairing samples with semantically aligned in-the-wild images, leveraging odor invariance within semantic categories. Whereas ScentFormer focuses on unimodal odor modeling, our framework adopts a two-stream transformer with a dedicated local alignment module to learn joint visuo-olfactory representations and dense cross-modal correspondences.

A concurrent unpublished work by Ozguroglu *et al.* [33] also studies visuo-olfactory learning, focusing on large-scale naturally paired data collection using a handheld sensing device to demonstrate visual supervision for odor representation learning. In contrast, we extend a smell-only dataset via synthetic pairing based on odor invariance. Architecturally, while their method employs global contrastive alignment, ours incorporates dense local alignment, enabling both fine-grained spatial correspondence and global representation learning. Additionally, we introduce smell localization as a new downstream task for visuo-olfactory grounding. Nevertheless, both works highlight the timeliness and importance of advancing visuo-olfactory multimodal learning.

3 Methodology

We aim to learn joint visuo-olfactory representations that generalize to downstream tasks including smell classification, cross-modal retrieval, and smell localization. To this end, we first construct a visuo-olfactory dataset by extending a smell-only dataset with semantically aligned in-the-wild web images. We then train a self-supervised framework that projects sniff and visual inputs into a shared embedding space and learns dense local cross-modal alignment via contrastive learning. An overview of the proposed framework is shown in Figure 2.

3.1 Construction of *SmellNet-V*

Learning joint visuo-olfactory representations requires paired multimodal training data. In the absence of such datasets, we construct a visuo-olfactory dataset, *SmellNet-V*, by synthetically pairing odor samples in the smell-only dataset [11] with semantically aligned in-the-wild web images. This pipeline produces multimodal data that approximate natural visuo-olfactory correspondences while increasing data granularity and enabling finer cross-modal alignment.

SmellNet Overview. SmellNet [11] is a large-scale dataset capturing real-world olfactory signals from 50 food and natural ingredients, grouped into nuts, spices, herbs, fruits, and vegetables, collected using portable multi-channel gas sensors. For each ingredient, 10 minutes of data were recorded per six sessions on different days, resulting in approximately 180,000 time-series samples at 1 Hz. Although 12 sensor channels are available, prior work primarily utilizes 6 chemically relevant components: NO₂, C₂H₅OH, VOC, CO, Alcohol and LPG. Importantly, SmellNet captures ingredients in their normal, non-degraded states.

Sniffing Units. In extending SmellNet to the visual domain, we draw inspiration from biological olfaction, where perception occurs through brief sniffing episodes rather than prolonged exposure [30,46]. Accordingly, we segment longer olfactory recordings into fixed-length temporal windows of size W with stride s , analogous to individual ‘*sniffs*’, and treat each sniff as a discrete training sample. This increases data granularity and aligns the training paradigm with natural olfactory sampling behavior.

Image Collection. To obtain diverse visual samples, we retrieve web images using context-rich queries for each ingredient. Rather than relying solely on class names, we use a large language model (LLM) [42] to generate descriptive phrases that capture varied real-world scenarios associated with each SmellNet ingredient. These queries are then used to collect visually diverse images (prompting details are provided in the suppl. material). For instance, for the ingredient ‘*apple*’, the LLM generates phrases such as ‘*whole apple on wooden table*’, ‘*apples in grocery produce section*’, and ‘*apple hanging on tree*’. By covering diverse environments and object configurations, the resulting image set captures broad visual variability within each ingredient category.

Image Filtering. After image collection, we apply a three-stage filtering pipeline. (1) *Prompt-based verification*: We use CLIP [34] with structured positive–negative prompt pairs to enforce (i) category consistency, (ii) photorealism (excluding drawings or illustrations), and (iii) valid object state (excluding spoiled or degraded instances). Images are retained only if their similarity to the positive prompt exceeds that of the corresponding negative prompts in all three tests. The exact prompt formulations are provided in the suppl. material. (2) *Quality filtering*: Images containing excessive watermarks (more than five detections using an off-the-shelf watermark detector [10]) are removed. (3) *Human refinement*: Finally, annotators perform lightweight manual verification to discard any remaining irrelevant samples.

Image-Sniff Pairing. After image collection and filtering, we construct visuo-olfactory training pairs by randomly matching each sniff unit with an image belonging to the same ingredient category. This category-level pairing yields discrete visuo-olfactory samples for training and forms the *SmellNet-V*.

Design Rationale of *SmellNet-V*. Our extension relies on the observation that odor identity is largely invariant to common visual transformations, including variations in lighting, scale, and appearance differences. For example, *a small red apple* and *a large red apple* emit nearly identical VOC signatures and are perceived under the same odor category, *apple*. This invariance allows us to pair any sniff unit within a ingredient category with visually diverse web images that share the same semantic identity. Moreover, because SmellNet captures ingredients in their normal, non-degraded states, visually matched web images in corresponding normal conditions are readily available, making such pairing physically consistent and semantically reliable. Thus, the resulting *SmellNet-V* approximates natural visuo-olfactory correspondences.

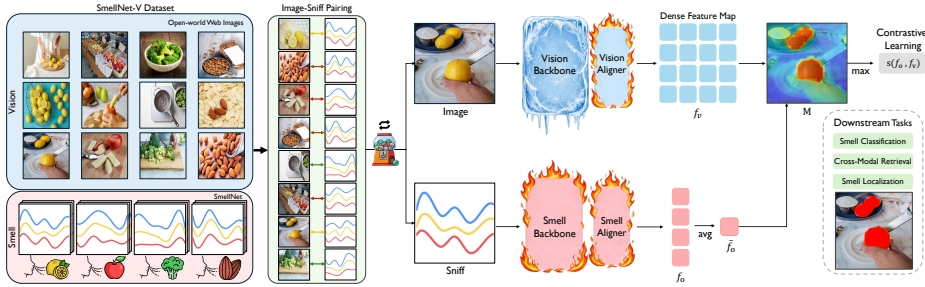


Fig. 2: Pipeline of See & Sniff. Our framework expands smell-only data through semantic pairing with web images. Vision and smell encoders extract modality-specific features, which are aligned using a contrastive objective to learn joint visuo-olfactory representations.

3.2 Model and Training Objective

Contrastive Learning aims to learn aligned representations by attracting positive pairs while repelling negative pairs. In the visuo-olfactory setting, let E_o and E_v denote the olfactory and visual encoders, respectively. For a sniff unit o_i , we obtain its feature $f_{o_i} = E_o(o_i)$, and for the corresponding visual sample v_i , we compute $f_{v_i} = E_v(v_i)$. The remaining visual features f_{v_j} for $j \neq i$, drawn from the dataset $\mathcal{D} = \{(v_i, o_i)\}_{i=1}^N$, serve as negatives. The objective \mathcal{L} is defined as:

$$\mathcal{L} = -\log \frac{\exp(s(f_{o_i}, f_{v_i})/\tau)}{\sum_j \exp(s(f_{o_i}, f_{v_j})/\tau)}, \quad (1)$$

where $s(\cdot, \cdot)$ denotes a cross-modal similarity function and τ is a temperature parameter [47]. Following prior studies [8, 14, 20, 34, 39, 48], we employ the loss in a symmetric manner across modalities.

Vision and Olfactory Encoders. Given an image v_i and its corresponding sniff unit o_i , the respective encoders extract modality-specific representations. Each encoder is composed of a backbone network followed by a lightweight aligner module. Through this pipeline, the raw inputs are projected into a shared representation space, producing a visual feature map $f_{v_i} \in \mathbb{R}^{C \times H \times W}$ and a sniff feature $f_{o_i} \in \mathbb{R}^{T \times C}$. Here, H and W denote the height and width of the feature map, T denotes the temporal sequence length of a sniff unit, and C indicates the channel dimension of the common embedding space.

Similarity Function. To enable fine-grained visuo-olfactory alignment, we adopt a similarity function that accounts for the characteristics of the task. Since our goal is to learn joint visuo-olfactory representations via dense local alignment, we first aggregate the sniff feature into a average pooled vector $\bar{f}_{o_i} = \frac{1}{T} \sum_{t=1}^T f_{o_i}[t]$, where $f_{o_i}[t] \in \mathbb{R}^C$ refers to the one-dimensional vector at the time step t of $f_{o_i} \in \mathbb{R}^{T \times C}$. We then compute a similarity map $M \in \mathbb{R}^{H \times W}$ by measuring the similarity between the spatial visual feature map and the aggregated sniff feature: $M[h, w] = \bar{f}_{o_i} \cdot f_{v_i}[h, w]$, where $f_{v_i}[h, w] \in \mathbb{R}^C$ denotes the one-dimensional feature vector at spatial location $[h, w]$, and \cdot indicates the inner product. We max-pool the similarity map to get the final similarity score $s(f_{o_i}, f_{v_i}) = \max(M)$.

3.3 Implementation Details

Model Architecture. Our framework consists of a visual encoder $E_v(\cdot)$ and an olfactory encoder $E_o(\cdot)$. For the visual branch, we adopt DINOv3-Small [41] pretrained weights and keep the backbone frozen during training. For the olfactory branch, we follow the ScentFormer from SmellNet [11], implemented as a 4-layer transformer [45] with 8 attention heads, trained from scratch. To project modality-specific features into a shared embedding space, we append lightweight aligners to each backbone. For the vision encoder, we use a channel-wise Layer-Norm [4] followed by a 1×1 convolution, which largely preserves the pretrained representation while adapting it to our objective. For the smell encoder, we employ a two-layer MLP with a residual connection. Both aligners are trained jointly under the proposed objective.

Training Pairs. The initial visuo-olfactory pairs in *SmellNet-V* are constructed via random ingredient-level matching between sniff units and web images as explained in Section 3.1. As the visuo-olfactory pairs are synthetically constructed rather than co-collected, we re-sample image–sniff pairs at each epoch to increase pairing diversity during training. Specifically, while the set of sniff units remains fixed, images within each ingredient are cyclically shifted. This strategy exposes each sniff unit to multiple visual instances across epochs, encouraging robust ingredient-level alignment rather than memorization of fixed synthetic pairs.

Training Setup. Our model takes a 224×224 image and a single sniff unit as input. For olfaction, we compute a first-order temporal difference with lag p , $\Delta x_t = x_t - x_{t-p}$, to emphasize relative changes in qualitative sensor outputs. We then extract windows of size (W, N_s) with stride $W/2$. The number of sensor channels N_s is 6, and we use $p=25$ and $W=50$ by default, unless specified otherwise. Training is performed on a NVIDIA RTX A5000 GPU with a batch size of 64. See suppl. for details.

4 Experiments

We evaluate our visuo-olfactory representations on *smell classification*, *cross-modal retrieval*, and *smell localization*. This section describes the datasets, baselines, and results.

4.1 Datasets

Training dataset. Our visuo-olfactory model is trained on *SmellNet-V*. With the default setting, *SmellNet-V* yields 5,411 sniff–image pairs for training.

Testing datasets. We evaluate our method on task-specific test sets as follows:

- **SmellNet-Test** [11]: The official test split of SmellNet, used for smell classification from smell alone. With the default setting, it yields 1,083 sniffs.
- ***SmellNet-V-Test***: A visuo-olfactory extension of the SmellNet test set constructed following our procedure in Section 3.1. It is used for cross-modal retrieval evaluation. The dataset contains the same number of smell samples as SmellNet-Test.

- **SmellNet-V-Source:** A newly constructed benchmark for smell source localization. Built upon *SmellNet-V-Test*, it includes manually annotated ingredient segmentation masks derived from the ground-truth ingredient categories. We use an interactive annotation tool [18] powered by SAM [24] to produce segmentation masks. Annotators provide sparse point prompts via simple mouse clicks to annotate regions according to predefined ingredient categories. Example annotations are in the second column of Figure 5.

4.2 Baselines

We compare our model with representative baselines, grouped by evaluation purpose, and specify the baselines used for each task.

Olfactory-Only vs. Visuo-Olfactory. To assess the impact of visual supervision on olfactory representation learning, we use ScentFormer from SmellNet as a unimodal baseline, evaluated only on ingredient smell classification task.

Global vs. Local Alignment. Self-supervised multimodal contrastive learning typically aligns modalities using inner products between global representations, such as pooled features or class tokens. While effective for tasks requiring holistic understanding (*e.g.*, linear probing or retrieval), global alignment may be insufficient for spatial localization. In contrast, dense local alignment based on similarity maps may enable both global correspondence and fine-grained grounding. The following baselines are included to compare global and local alignment strategies: (1) *See & Sniff*: Our proposed model trained with the dense local alignment objective, (2) *Ours-Local*: Local alignment model without aligners for each modality encoder, (3) *Ours-Global*: A variant of *Ours-Local* using a CLS-token based global alignment objective, (4) *Global-CLIP*: A baseline combining a frozen CLIP-Large [34] pre-trained visual encoder with our olfactory encoder, optimized using a CLIP-style global contrastive objective. The global baselines also approximate the recent work [35]. These baselines are used for all tasks.

Below baselines are only used for smell localization:

Visual Bias Analysis. To examine potential visual biases in the localization benchmark, we include several vision-only baselines that do not rely on olfactory input. (1) *Full Square* and (2) *Full Circle*: fixed binary masks (a 224×224 square or a circle with diameter 224) applied uniformly without any visual or smell understanding. (3) *DINOv3 Attention Map*: a vision-only baseline derived from attention maps of a pre-trained DINOv3 model, capturing generic objectness cues without smell information.

Cascaded Approach. We consider a two-stage pipeline, SmellNet + SAM3, that first predicts smell categories using SmellNet and subsequently performs text-conditioned segmentation (*e.g.*, SAM3 [5]) for localization. This cascaded approach serves as a baseline to compare against our *See & Sniff* model.

Upper Bound Baseline. We also include an upper-bound reference based on the cascaded approach. Here, ground-truth ingredient category labels are fed to SAM3 to get segmentation masks. This baseline estimates the maximum achievable localization when the smell classification is assumed to be perfect.

Model	$p = 25, W = 50$		$p = 25, W = 100$	
	Acc.	F1	Acc.	F1
SmellNet (LSTM)	50.6	48.8	57.9	56.0
SmellNet (Transformer)	50.6	49.5	56.1	55.5
Global-CLIP	53.19	52.22	61.55	60.01
Ours-Global	53.74	52.64	59.36	57.96
Ours-Local	<u>54.94</u>	<u>53.78</u>	<u>62.75</u>	<u>62.31</u>
See & Sniff	57.71	56.68	63.75	62.66

Table 1: Smell Classification.

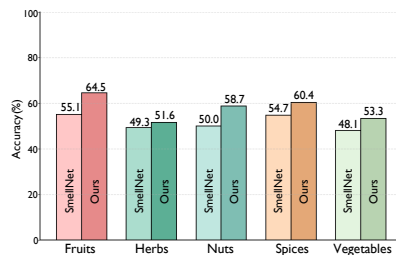


Fig. 3: Family-wise Comparison.

4.3 Main Results

– Smell Classification Task

We evaluate the quality of learned olfactory representations on ingredient classification across the 50 categories defined in SmellNet. After self-supervised training of *See & Sniff* on *SmellNet-V*, we freeze the olfactory encoder and train a linear probe using unimodal olfactory signals with the labels from SmellNet. Evaluation is conducted on the smell-only SmellNet-Test split. Results are in Table 1.

Key findings are as follows:

- (1) *Visual supervision via SmellNet-V and See & Sniff learns stronger olfactory representations.* Across all settings and variants, *See & Sniff* trained on *SmellNet-V* surpasses smell-only SmellNet baselines under smell-only inference, demonstrating that vision provides effective supervision for olfactory representation learning and validating our synthetic visuo-olfactory pairing strategy.
- (2) *The gains are robust across sniff configurations.* We report results under SmellNet’s original configurations ($p=25, W=50$ and $p=25, W=100$) for fair comparison. Under all settings, our model consistently outperforms SmellNet baselines, showing that performance gains are not tied to a specific parameter choice.
- (3) *Dense local alignment with aligner yields the best performance.* While global alignment provides competitive performance for holistic understanding, dense local alignment consistently achieves superior results. Furthermore, incorporating the lightweight aligner yields additional gains, supporting our architectural design for effective unimodal and cross-modal representation learning.

Additional Analysis. We further analyze smell classification across the five major ingredient families in SmellNet, including fruits, herbs, nuts, spices, and vegetables. Specifically, we report averaged ingredient-level classification results within each family to examine whether the observed gains are consistent across families rather than driven by a specific group. As shown in Figure 3, our model consistently improves performance across all five families compared to the smell-only baseline. Notable gains are observed for fruits, nuts, vegetables, and spices, with a modest but consistent improvement for herbs. The larger improvements in fruits, nuts, and vegetables can be attributed to their distinctive object-level visual structures, such as clear shape, color, and geometry, which provide informative cross-modal supervision when paired with olfactory signals. In contrast, herbs often exhibit fine-grained leaf textures and visually similar appearances across ingredients, limiting the additional discriminative cues available from vi-

Model	Smell \rightarrow Vision			Vision \rightarrow Smell		
	R@1	R@5	R@10	R@1	R@5	R@10
Global-CLIP	53.28	55.96	58.82	<u>56.90</u>	87.40	<u>91.40</u>
Ours-Global	50.97	62.60	67.50	52.80	78.30	85.30
Ours-Local	<u>53.65</u>	58.45	61.50	55.40	82.90	88.90
See & Sniff	56.14	<u>60.94</u>	<u>63.90</u>	63.20	87.40	91.90

Table 2: Cross-modal Retrieval.



Fig. 4: Smell \rightarrow Vision Retrieval. The top-10 images retrieved by given smell queries are shown. Blue borders indicate correct matches, and red borders indicate mismatches. The failure cases are visually similar to the queried ingredients, indicating fine-grained visual ambiguity rather than random mismatch.

sual supervision. Nevertheless, the consistent improvement across all ingredient families suggests that visuo-olfactory pairing enhances olfactory representation learning in a broad and family-agnostic manner.

– Cross-Modal Retrieval Task

We evaluate cross-modal retrieval to measure the alignment between visual and olfactory embeddings. Given a sniff, the task is to retrieve the image of the same ingredient, and vice versa. Retrieval results are obtained by ranking the cosine similarity between the query embedding and embeddings from the other modality. Results are reported on the *SmellNet-V* test set using Recall@K as the metric and are shown in Table 2.

Quantitative Results. The retrieval results demonstrate that *SmellNet-V* provides a viable foundation for visuo-olfactory training, as all variants exhibit cross-modal understanding, and that *See & Sniff* effectively learns cross-modal correspondences. It achieves the best overall performance in 4 of 6 metrics. As expected, global alignment performs strongly in retrieval, particularly on metrics such as R@5 and R@10 for Smell \rightarrow Vision; however, our final model consistently outperforms it in the more challenging R@1 setting for both directions, indicating stronger fine-grained matching between visual and olfactory embeddings. Incorporating the aligner further improves performance, suggesting a more stable shared representation space that benefits cross-modal retrieval. Together with the improvements observed in smell classification and cross-modal retrieval, these results validate both our dataset construction strategy and alignment design.

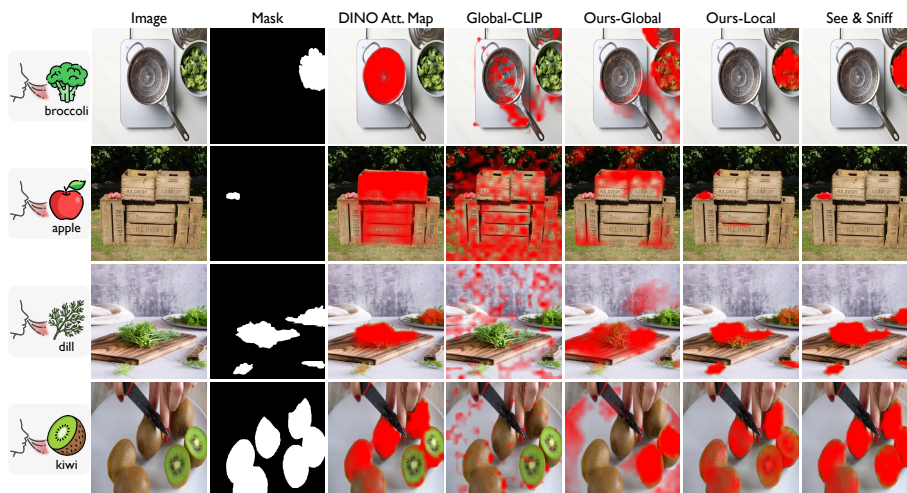


Fig. 5: Qualitative Smell Localization Results.

Qualitative Results. The results are in Figure 4. The model retrieves semantically consistent images conditioned on the input smell. For instance, the strawberry smell retrieves images of strawberries across diverse contexts, including sliced fruits, topping on the cake, and natural scenes. Similarly, the asparagus smell retrieves relevant asparagus images. A failure case (marked in red) retrieves chives, which are visually similar due to their elongated green structure. This error reflects fine-grained visual ambiguity rather than random mismatch.

Smell→Smell Retrieval. We further evaluate Smell→Smell retrieval (in Table 3) to assess the intrinsic structure of the learned olfactory embeddings. Compared to the SmellNet baseline, our model achieves superior retrieval performance, indicating that visuo-olfactory training enhances the semantic structure of the smell embedding space. Since retrieval relies solely on smell features at inference time, these gains suggest that visual supervision also regularizes and strengthens intra-modal representations rather than simply enabling cross-modal matching.

Method	R@1	R@5	R@10
SmellNet	52.54	76.18	85.50
See & Sniff	69.44	85.04	89.47

Table 3: Smell→Smell Retrieval.

– Smell Localization Task

We introduce smell localization as a new task in machine olfaction, aiming to spatially ground a given sniff within a visual scene. Given an olfactory input, the objective is to localize the image region expected to emit the corresponding smell. *SmellNet-V-Source* is used as test set for this task. We use mAP and mIoU as evaluation metrics by following standard multimodal grounding protocols [7, 9, 16, 28, 36]. Results are in Table 4 and Figure 5. **Key findings** are as follows:

Method	mAP	mIoU
<i>Binary mask</i>		
Full square	–	0.2072
Full circle	–	0.2393
<i>Visual heatmap</i>		
DINOv3 Att. Map	0.6596	0.5157
<i>Global alignment</i>		
Global-CLIP	0.1736	0.2073
Ours-Global	0.6676	0.5016
<i>Local alignment</i>		
Ours-Local	0.7970	0.6099
See & Sniff	0.8362	0.6456
<i>Cascaded</i>		
SmellNet + SAM3	–	0.3214
SmellNet + SAM3 [†]	–	0.3683
GT + SAM3	–	0.5854
GT + SAM3 [†]	–	0.6700

Table 4: Smell Localization Results.

spondences during representation learning. These results demonstrate that joint end-to-end multimodal learning is more effective than post-hoc category-conditioned segmentation.

(3) *Competitive performance against the GT upper-bound baseline.* While GT + SAM3 serves as an upper-bound reference by using ground-truth category labels, SAM3 exhibits inherent limitations for certain smell categories due to out-of-domain effects. To ensure fairness, we additionally report results excluding 7 categories where SAM3 fails to produce meaningful segmentations (marked with †). Even under this favorable setting, our model remains competitive and surpasses it without this adjustment, demonstrating the robustness of the learned visuo-olfactory representations.

(4) *Visuo-olfactory alignment is necessary for smell localization.* Vision-only baselines such as DINOv3 Att. Maps fail to match our model, indicating that generic visual objectness is insufficient for smell localization. Accurate grounding requires explicit visuo-olfactory alignment rather than purely visual cues.

(5) *The aligner module further enhances cross-modal alignment.* Incorporating the aligner improves the performance of the local alignment variant (Ours-Local vs. See & Sniff). This improvement suggests that projecting modality-specific features into a common embedding space strengthens visuo-olfactory correspondence, leading to better smell localization.

Qualitative Results. The results are shown in Figure 5. *See & Sniff* accurately localizes the target source even when it is off-center, or when multiple source objects are present. For example in the second row, while other baselines focus on the visually salient wooden boxes, our local alignment models correctly identify the apple on the left.

(1) *Local alignment is essential for spatial smell grounding.* Local alignment substantially outperforms global alignment in smell localization. While global embeddings capture holistic correspondence, they fail to model spatial cross-modal interactions. In contrast, our dense local alignment objective learns fine-grained visuo-olfactory correspondences, enabling accurate smell source grounding. This directly validates our architectural design and learning objective.

(2) *Joint multimodal learning outperforms cascaded pipelines.* The cascaded baseline (SmellNet + SAM3) performs significantly worse than our unified model. This two-stage pipeline lacks direct visuo-olfactory feature interaction and cannot model spatial corre-

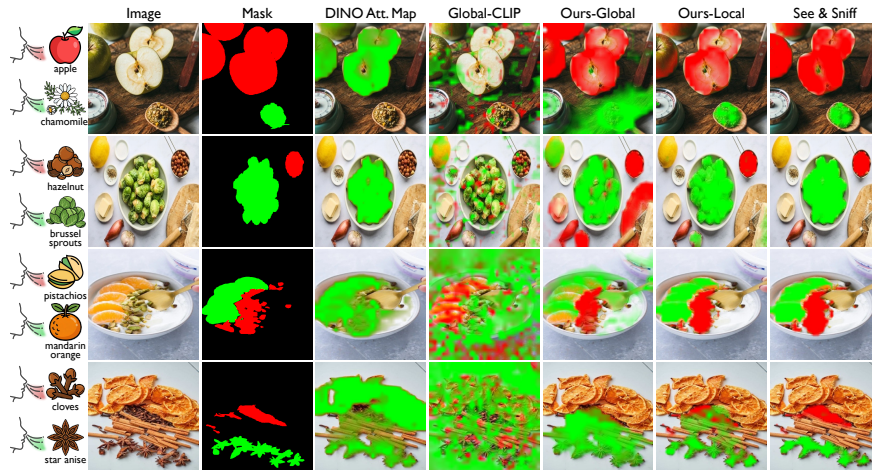


Fig. 6: Qualitative Results on Interactive Localization.

Interactive Localization. To further examine fine-grained visuo-olfactory alignment, we evaluate our model under an interactive localization setting following [40]. In this setup, a single image containing multiple ingredients is paired with different smell inputs, and a reliable localization method should shift its predicted region according to the given smell. This analysis verifies that localization is driven by smell semantics rather than fixed visual saliency.

We construct a new dataset, *SmellNet-V-InteractiveSource*, for interactive smell localization. It contains 100 images, each depicting two distinct ingredients with pixel-level segmentation masks. Forty images are drawn from the *SmellNet-V-Source* test set, while the remaining are newly collected from the web to ensure the presence of multiple ingredients within a single scene. For evaluation, each smell signal is divided into multiple sniff units as before. For a given image, localization performance is computed independently for each sniff unit corresponding to each ingredient, and the IoU scores are averaged to obtain a final score per ingredient. A sample is considered successful only if the IoU for both ingredients exceeds 0.5, ensuring that the model correctly localizes each smell source within the same scene. As shown in Table 5, interactive localization further shows the importance of local cross-modal alignment. Global-alignment methods perform poorly, with Global-CLIP failing entirely and Ours-Global achieving only limited success. DINOv3 attention maps also struggle, as they inherently produce a single dominant region.

Model	IIoU
<i>Visual heatmap</i>	
DINOv3 Att. Map	0.01
<i>Global alignment</i>	
Global-CLIP	0.00
Ours-Global	0.22
<i>Local alignment</i>	
Ours-Local	<u>0.32</u>
See & Sniff	0.35
<i>Cascaded</i>	
SmellNet + SAM3	0.06
GT + SAM3	0.41

Table 5: Interactive Localization Results.

In contrast, *See & Sniff* outperforms all baselines except the upper-bound setting. These findings indicate that interactive smell grounding requires fine-grained, smell-conditioned spatial reasoning rather than static saliency or global embedding similarity. Although the overall IoU values reflect the difficulty of the task, our model consistently shows superior conditional grounding, as further illustrated in Figure 6.

Other Discussion. Lastly, we analyze localization under different physical states of the same ingredient, for example, peanuts with varying shell conditions; peeled, sliced, or crushed garlic; whole or sliced cabbage with varying colors; or ground versus non-ground cinnamon. Qualitative results are shown in Figure 7. Our model successfully localizes peanuts, garlic and cabbage across diverse visual appearances within the same scene, demonstrating robustness to state-level variations. However, we observe failure cases for powdered spices, as the nearly identical brown, fine-grained textures of different ingredients within the family limit discriminative cues for localization. These findings suggest that while the model is robust to structural variations, texture-dominated states remain challenging due to limited visual separability.



Fig. 7: Localization across Physical States.

5 Conclusion and Discussion

In this work, we take an early step toward visuo-olfactory learning by transforming a unimodal smell dataset into a cross-modal training paradigm through synthetic pairing with in-the-wild images, forming *SmellNet-V*, and by introducing *See & Sniff*, a self-supervised model with dense local alignment for joint representation learning. Our results show that visual supervision not only enables cross-modal retrieval and smell localization, but also strengthens intrinsic olfactory representations, surpassing smell-only learning by a significant margin in classification. By establishing new benchmarks for classification, retrieval, and spatial grounding, we move machine olfaction beyond unimodal modeling and open a path toward integrating smell into future multimodal perception systems.

What can be further done? Future work may extend visuo-olfactory learning beyond normal, non-degraded ingredient states to scenarios involving spoilage or transformation, where visual and olfactory cues encode temporal semantic changes (*e.g.*, fermentation or decay). Another promising direction is modeling compositional and mixture odors, enabling recognition and localization of complex food items or multi-ingredient scenes (*e.g.*, lemon cheesecake conditioned on lemon scent). Such extensions would move toward more realistic and semantically rich visuo-olfactory perception.

6 Acknowledgment

This work was supported by IITP grants funded by the Korean government (MSIT) (RS-2024-00457882, National AI Research Lab Project, 50%; RS-2020-II201336, Artificial Intelligence Graduate School Program, UNIST, 10%); the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (RS-2026-25496684, 35%); and the Hankuk University of Foreign Studies Research Fund (of 2026, 5%).

References

1. Achebouche, R., Tromelin, A., Audouze, K., Taboureau, O.: Application of artificial intelligence to decode the relationships between smell, olfactory receptors and small molecules. *Scientific reports* (2022)
2. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: *ICCV* (2017)
3. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: *NeurIPS* (2016)
4. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. *arXiv preprint arXiv:1607.06450* (2016)
5. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. In: *ICLR* (2026)
6. Castellotti, S., Soldo, M., Plank, T., Viva, M.M.D., Greenlee, M.W.: Visual search performance depends on the congruency of olfactory sensations. *Scientific Reports* (2025)
7. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: *CVPR* (2022)
8. Elizalde, B., Deshmukh, S., Ismail, M.A., Wang, H.: Clap: Learning audio concepts from natural language supervision. In: *ICASSP* (2023)
9. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. *IJCV* (2015)
10. Fancy Feast: Joycaption Watermark Detection. *Hugging Face Spaces* (2025), <https://huggingface.co/spaces/fancyfeast/joycaption-watermark-detection>, Accessed 24 June 2026
11. Feng, D., Dai, W., Li, C., Pernigo, A., Wen, Y., Liang, P.P.: Smellnet: A large-scale dataset for real-world smell recognition. In: *ICLR* (2026)
12. Fichtelmann, P., Westermayr, J.: Machine learning for smell: Ordinal odor strength prediction of molecular perfumery components. *arXiv preprint arXiv:2512.08683* (2025)
13. Fu, L., Datta, G., Huang, H., Panitch, W.C.H., Drake, J., Ortiz, J., Mukadam, M., Lambeta, M., Calandra, R., Goldberg, K.: A touch, vision, and language dataset for multimodal alignment. In: *ICML* (2024)
14. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: *CVPR* (2023)
15. Gong, Y., Rouditchenko, A., Liu, A.H., Harwath, D., Karlinsky, L., Kuehne, H., Glass, J.: Contrastive audio-visual masked autoencoder. In: *ICLR* (2022)
16. Hamilton, M., Zisserman, A., Hershey, J.R., Freeman, W.T.: Separating the "chirp" from the "chat": Self-supervised visual grounding of sound and language. In: *CVPR* (2024)

17. Harwath, D., Recasens, A., Surís, D., Chuang, G., Torralba, A., Glass, J.: Jointly discovering visual objects and spoken words from raw sensory input. In: ECCV (2018)
18. Intel Corporation: CVAT: Computer Vision Annotation Tool (2025), <https://www.cvat.ai/>, Accessed 24 June 2026
19. Iwata, H.: Interpretable multitask deep learning models for odor perception based on molecular structure. *Current Research in Food Science* (2025)
20. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML (2021)
21. Kang, D., Kim, J., Park, J., Lee, K., Choi, J.W., So, J.: Aromma: Unifying olfactory embeddings for single molecules and mixtures. *arXiv preprint arXiv:2601.19561* (2026)
22. Keller, H.: A neglected treasure. *The Home Magazine* (1934), <https://www.afb.org/HelenKellerArchive?a=d&d=A-HK02-B225-F02-024>
23. Kim, S., Lee, S., Ryu, H., Chung, J.S., Senocak, A.: Seeing through touch: Tactile-driven visual localization of material regions. In: CVPR (2026)
24. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV (2023)
25. Lee, B.K., Mayhew, E.J., Sanchez-Lengeling, B., Wei, J.N., Qian, W.W., Little, K.A., Andres, M., Nguyen, B.B., Moloy, T., Yasonik, J., et al.: A principal odor map unifies diverse tasks in olfactory perception. *Science* (2023)
26. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023)
27. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML (2022)
28. Lüddecke, T., Ecker, A.: Image segmentation using text and image prompts. In: CVPR (2022)
29. Lyu, Y., Zheng, X., Kim, D., Wang, L.: Omnibind: Teach to build unequal-scale modality interaction for omni-bind of all. *arXiv preprint arXiv:2405.16108* (2024)
30. Mainland, J., Sobel, N.: The sniff is part of the olfactory percept. *Chemical senses* (2006)
31. Mueller, P., Salminen, K., Nieminen, V., Kontunen, A., Karjalainen, M., Isokoski, P., Rantala, J., Savia, M., Väliäho, J., Kallio, P., et al.: Scent classification by k nearest neighbors using ion-mobility spectrometry measurements. *Expert systems with applications* (2019)
32. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: ECCV (2018)
33. Ozguroglu, E., Liang, J., Liu, R., Chiquier, M., DeTienne, M., Qian, W.W., Horowitz, A., Owens, A., Vondrick, C.: New york smells: A large multimodal dataset for olfaction. *arXiv preprint arXiv:2511.20544* (2025)
34. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
35. Raitzel, C.U., Gottfried, J.A.: Using your nose to find your way: Ethological comparisons between human and non-human species. *Neuroscience & Biobehavioral Reviews* (2021)
36. Ryu, H., Kim, S., Chung, J.S., Senocak, A.: Seeing speech and sound: Distinguishing and locating audio sources in visual scenes. In: CVPR (2025)

37. Sanchez-Lengeling, B., Wei, J.N., Lee, B.K., Gerkin, R.C., Aspuru-Guzik, A., Wiltschko, A.B.: Machine learning for scent: Learning generalizable perceptual representations of small molecules. arXiv preprint arXiv:1910.10685 (2019)
38. Senocak, A., Oh, T.H., Kim, J., Yang, M.H., Kweon, I.S.: Learning to localize sound source in visual scenes. In: CVPR (2018)
39. Senocak, A., Ryu, H., Kim, J., Oh, T.H., Pfister, H., Chung, J.S.: Sound source localization is all about cross-modal alignment. In: ICCV (2023)
40. Senocak, A., Ryu, H., Kim, J., Oh, T.H., Pfister, H., Chung, J.S.: Toward interactive sound source localization: Better align sight and sound! IEEE TPAMI (2025)
41. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
42. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
43. Sung, S.H., Suh, J.M., Hwang, Y.J., Jang, H.W., Park, J.G., Jun, S.C.: Data-centric artificial olfactory system based on the eigengraph. Nature communications (2024)
44. Tran, N., Kepple, D., Shuvaev, S., Koulakov, A.: Deepnose: Using artificial neural networks to represent the space of odorants. In: ICML (2019)
45. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
46. Wachowiak, M.: All in a sniff: olfaction as a model for active sensing. Neuron (2011)
47. Wu, Z., Xiong, Y., Yu, S., Lin, D.: Unsupervised feature learning via non-parametric instance-level discrimination. In: CVPR (2018)
48. Yang, F., Feng, C., Chen, Z., Park, H., Wang, D., Dou, Y., Zeng, Z., Chen, X., Gangopadhyay, R., Owens, A., et al.: Binding touch to everything: Learning unified multimodal tactile representations. In: CVPR (2024)
49. Yang, F., Ma, C., Zhang, J., Zhu, J., Yuan, W., Owens, A.: Touch and go: Learning from human-collected vision and touch. In: NeurIPS - Datasets and Benchmarks Track (2022)
50. Yang, F., Zhang, J., Owens, A.: Generating visual scenes from touch. In: ICCV (2023)
51. Zhai, X., Mustafa, B., Kolesnikov, A., Beyler, L.: Sigmoid loss for language image pre-training. In: ICCV (2023)