What Images Cannot Say: Language-Guided Olfactory Representation Learning

Eleftherios Tsonis, Xi Wang, and Vicky Kalogeiton

LIX, École Polytechnique, IP Paris, CNRS

{firstname.lastname}@polytechnique.edu

https://www.lix.polytechnique.fr/vista/projects/2026_scent_tsonis/

Abstract. Images tell us what a scene looks like, but rarely what it would feel like to be there. While recent datasets pair visual scenes with electronic-nose measurements, aligning smell signals with images remains challenging because many olfactory cues arise from contextual environmental factors that are not directly visible in pixels. We introduce SCENT, a multimodal framework that uses language guidance as a semantic bridge between vision and olfaction. Our approach leverages Vision-Language Models (VLMs) to generate scene descriptors capturing objects, environmental context, and plausible ambient smell cues suggested by the visual scene. These descriptors provide semantic guidance for learning olfactory representations. We train a smell encoder that maps electronic-nose signals into a shared embedding space aligned with both visual and textual representations, and introduce a language-guided latent decomposition that separates object-specific odors from contextual environmental contributions. Experiments on the New York Smells dataset demonstrate that SCENT significantly improves cross-modal retrieval compared to vision-only baselines, achieving state-of-the-art performance on smell-to-image and smell-to-text retrieval tasks. In addition, our framework produces interpretable olfactory representations that enable the disentanglement of complex smell mixtures. Our results reveal the importance of contextual semantic information for grounding olfactory perception in multimodal learning and pave the way for future research in this area.

Keywords: Cross-modal Retrieval · Olfactory Representation Learning · Vision-Language Models

1 Introduction

"A picture is worth a thousand words" (att. Frederick R. Barnard)
*... but it rarely tells us what it **feels** like to be there..*

In real environments, perception extends beyond vision. A photograph of a busy street may evoke the smell of vehicle exhaust; a metro entrance may imply the metallic scent of ventilation air. Although such environmental cues are rarely

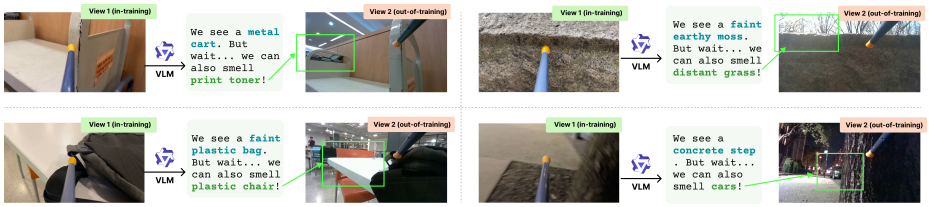


Fig. 1: Not everything we can smell is visible. Given only View 1 (in-sample) during training, the true olfactory context may lie outside the field of view, in View 2 (out-of-sample). A VLM can bridge this gap by inferring plausible smells from semantic context alone. We use these language-derived signals as supervision to learn richer smell representations that go beyond what is directly seen.

visible directly, humans routinely infer them from visual context. Despite years of progress in computer vision, AI systems today primarily capture *what a scene looks like*, while remaining largely unaware of what it might *feel like to be there*.

Recent advances in Vision-Language Models (VLMs) suggest that this long-standing idea may finally be realized: modern systems can generate rich descriptions of visual scenes [33, 34, 65], answer questions about images [18, 62, 69], and reason about complex environments through language [5, 29]. Their extensive world knowledge opens the possibility of inferring what exists in a scene, beyond what is visible.

Among sensory modalities, olfaction is particularly important. Smell conveys rich semantic information that is difficult to infer from other modalities [21, 52], including food quality and safety [1, 53], environmental conditions and airborne hazards, such as pollution [6] and toxic substances [7, 61], as well as social cues related to health and emotional state [30, 39].

Advances in electronic noses (e-noses) [61] enable the capture of high dimensional sensor measurements [17,19,44] that characterize the chemical composition of air. Mapping these raw signals to semantic representations is, however, fundamentally difficult. First, the image paired with each recording captures only a *partial view* of the scene: smells diffuse freely and may originate from sources entirely outside the camera's field of view, so visual supervision alone cannot fully account for what was measured (Figure 1). Second, a smell recording reflects a mixture of contributions from multiple odor sources in the scene, making it hard to disentangle individual semantic factors.

VLMs trained on large-scale data develop world knowledge that extends well beyond visual appearence, learning to reason about physical, acoustic, and semantic properties of scenes [26, 58, 67]. This extends to olfaction, as Large Language Models (LLM) can classify odors, predict smell descriptors, and identify sources from natural language descriptions, demonstrating that smell associations are encoded through semantic context [38]. Inherited from LLMs, modern VLMs also hold natural knowledge priors which are useful for our problem: given a scene image, a VLM can generate plausible olfactory descriptions that

go beyond what the camera captures, directly compensating for the partial observability of visual supervision.

Building on this insight, we propose **SCENT**: **S**emantic **C**ontext-aware **e-N**ose **T**ransformer, a framework that integrates e-nose measurements with visual and linguistic representations. Our approach first uses a VLM to generate structured textual descriptors of a scene, capturing objects, environmental context, and plausible ambient smell cues. These descriptions serve as a semantic bridge between vision and smell: they encode what a scene implies about its olfactory environment, extending supervision beyond what the camera alone can observe. The descriptors are encoded and used as semantic anchors for representation learning. We train a smell encoder that maps e-nose signals into a shared embedding space aligned with both visual and textual representations, enabling multimodal representation of olfactory environments.

To further structure the learned representation, we introduce a latent decomposition that separates object-specific odor signals from contextual environmental contributions. This decomposition is guided by the text descriptions produced by the VLM and encourages the model to disentangle the components of complex olfactory mixtures.

We evaluate our approach on the New York Smells (NYS) dataset [44] and demonstrate that integrating language guidance significantly improves multimodal alignment compared to vision-only baselines. Our method achieves state-of-the-art performance on cross-modal retrieval tasks, including smell-to-image and smell-to-text retrieval, while also producing more interpretable olfactory representations. In summary, our contributions are:

- We introduce the idea of using **language guidance as a semantic bridge** between visual scenes and olfactory signals, enabling the interpretation of smell measurements through contextual world knowledge.
- We present SCENT, a **multimodal framework for learning olfactory representations** that aligns e-nose, visual, and textual embeddings.
- We propose a language-guided **decomposition of olfactory representations** that separates object-related odors from environmental contributions.

2 Related Work

Cross-modal and multimodal supervision. Using one modality to supervise another has a long history, spanning early self-supervised approaches [11] and deep multimodal models [41], to more recent methods that exploit naturally co-occurring sensor data for audio-visual [2, 43] and visual-tactile [13, 66, 68] learning. Models such as CLIP [45] and ALIGN [27] learn joint image-text representations, while CLAP [16] extends this paradigm to audio and text. Similar approaches have since been applied to video–text [37] and point cloud-text [70] representation learning. Works such as CLIP4VLA [48], VALOR [35], VAST [8], mPLUG-2 [63], UMT [36], InternVideo2 [59], and ConFu [31] align three modalities at once. For higher order settings, existing methods often designate one modality as an anchor and align the remaining modalities through pairwise

contrastive objectives. For example, ImageBind [23] uses images as the central bridge, and LanguageBind [73] uses text. OneLLM [25] aligns all modalities to a frozen language model, and UniAlign [72] encodes modalities in a unified mixture-of-experts architecture. More recently, alternative formulations have been explored, including geometric approaches such as GRAM [10] and TRIANGLE [9], and information-theoretic objectives such as Symile [50] and CoMM [15].

Language as a semantic bridge across modalities Machine generated language provides a scalable alternative to manual annotation, enabling semantic supervision of modalities that are otherwise expensive to label. Early neural image captioning [56] established that vision and language can be jointly modelled to produce descriptive text from visual inputs. Later work leverages this ability as a supervisory signal across diverse modalities. ULIP-2 [64] generates holistic language descriptions from rendered views of 3D shapes, enabling scalable trimodal pre-training over point clouds, images, and text without any human annotation. LAVILA [71] similarly uses LLMs as narrators to densely annotate video, yielding supervision for contrastive video–text learning. Moving beyond vision, WineSensed/FEAST [4] demonstrates that text and images can be aligned to human sensory perception of taste, establishing that sensory modalities can be grounded in shared multimodal representations. Building on these precedents, we use VLM-generated descriptors to inject world knowledge and contextual cues into olfactory representation learning.

Machine olfaction and olfactory representations. Machine olfaction has historically focused on constrained settings using specialized or laboratory-grade sensing [12, 19, 40, 47, 55, 60], enabling applications such as scent design [49], disease detection [20, 22], and security [54]. At the molecular level, prior work has used psychophysical data and graph-based models to predict perceptual attributes [14, 28] and to define a principal odor map (POM) [32], while mixture and similarity studies remain limited [46, 51]. Recent findings from neuroscience emphasize the importance of studying olfaction under natural concentration ranges [57]. Concurrently, recent work examines how well large language models can reason about smell [38].

Moving beyond lab-based limitations, the New York Smells (NYS) dataset [44] is a large-scale multimodal benchmark designed for "in-the-wild" smell perception. It comprises 7,000 image-olfactory pairs, collected across diverse urban environments ranging from indoor libraries to outdoor parks. NYS frames olfactory learning as direct cross-modal alignment between visual imagery and e-nose signals. Our approach differs from NYS by using language as a semantic bridge to inject *prior world knowledge* into smell representation learning, capturing olfactory cues that visual appearence alone cannot express. To our knowledge, no existing work combines *olfactory sensor signals* with *vision* and *text* together.

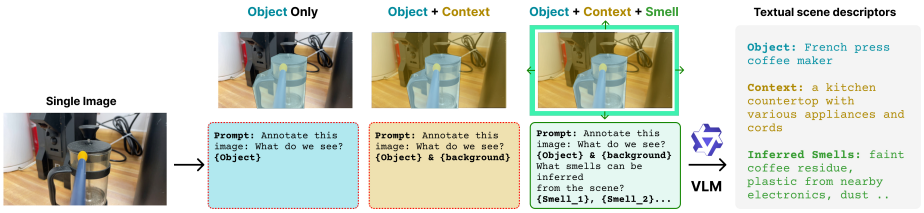


Fig. 2: Textual scene descriptors via VLM inference. Given an image, a pretrained VLM generates object, contextual, and inferred smell descriptors, which serve as language guidance for olfactory representation learning.

3 Method

We introduce **SCENT**: **S**emantic **C**ontext-aware **e-N**ose **T**ransformer, a framework for learning olfactory representations from e-nose measurements, guided by vision and language. Our key insight is that an olfactory signal $\mathbf{X} \in \mathbb{R}^{C \times T}$ (a C-channel e-nose recording over T time steps) reflects odor contributions from the *entire* scene, yet paired visual observations capture only a subset of these factors (Figure 1). In other words, smell information is only **partially observable** from visual input. To bridge this gap, we leverage the strong world priors of Vision-Language Models (VLMs) [33, 65], which can infer plausible olfactory cues from visual context.

SCENT therefore combines three stages: (1) VLM-based semantic scene augmentation (Figure 2, Section 3.1); (2) multimodal olfactory representation learning (Figure 3(a), Section 3.2); and (3) latent smell disentanglement that separates object and contextual odor components (Figure 3(b), Section 3.3).

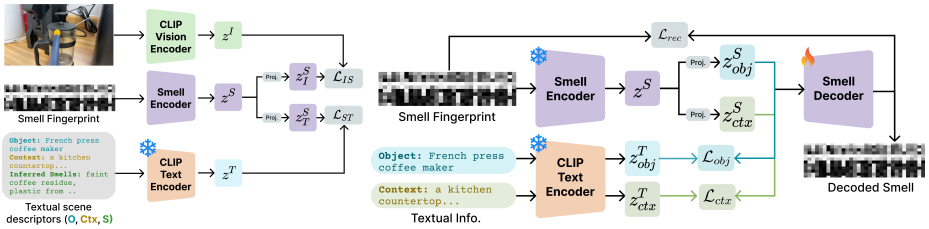
3.1 VLM-based Semantic Scene Augmentation

The NYS dataset provides images I and corresponding smell measurements $\mathbf{X}$. To expose the environmental factors that shape olfactory perception, we query a pretrained VLM through a structured prompt, eliciting its prior world knowledge at three semantic levels. Given an image I, the VLM produces a set of language descriptors:

$$T(I) = \{t_{\text{obj}}, t_{\text{ctx}}, t_{\text{smell}}^{(1)}, \ldots, t_{\text{smell}}^{(K)}\} \quad ,$$

where t_{obj} denotes a textual description of the primary object in the scene, t_{ctx} describes the surrounding environmental context, and $\{t_{\text{smell}}^{(k)}\}_{k=1}^{K}$ are textual descriptors referring to plausible ambient smell cues suggested by the scene context. These descriptors provide semantic cues that may relate to the measured olfactory signal but are not directly encoded in the image alone. As shown in Figure 2, we explore several prompting schemes and adopt the rightmost one, which explicitly encourages the VLM to infer potential olfactory elements.

These descriptors are combined and prompted into a textual description and encoded using the frozen CLIP [45] text encoder: $z^T = f_T(\text{prompt}(T)) \in \mathbb{R}^{512}$.



(a) Smell representation learning. **(b)** Latent smell disentanglement.

Fig. 3: Overall framework for multimodal olfactory representation learning and disentanglement. (a) Smell Representation Learning: We train a Smell Encoder to align smell fingerprints with CLIP visual (z^I) and textual (z^T) embeddings. The dual projection heads yield z^S_I and z^S_T. **(b)** Latent Smell Disentanglement: Leveraging the pretrained encoder, we disentangle the smell representation into object (z^S_{obj}) and context (z^S_{ctx}) components. These components are aligned with their respective linguistic descriptors via $\mathcal{L}_{\text{obj}}$ and $\mathcal{L}_{\text{ctx}}$.

The resulting representation z^T serves as language supervision in the multimodal alignment stage (Section 3.3), helping compensate for information missing from visual observations.

3.2 Learning Multi-modal Olfactory Representations

With visual observations I and augmented smell descriptors $T(I)$ available, we learn an olfactory representation aligned with both visual and textual modalities.

Olfactory Encoder. The smell signal $\mathbf{X}$ is processed using a Transformer-based encoder f_S that models temporal dependencies across sensor measurements, yielding $z^S = f_S(\mathbf{X})$. This embedding captures the global structure of the olfactory signal. To enable alignment with visual and textual modalities, we apply two modality-specific projection heads:

$$z^S_I = \phi_I(z^S) \in \mathbb{R}^{512}, \qquad z^S_T = \phi_T(z^S) \in \mathbb{R}^{512},$$

which map the olfactory representation toward the visual and textual embedding spaces, respectively (Figure 3a).

Visual and Textual Encoders. We obtain visual embeddings z^I using CLIP's [45] image encoder, and textual embeddings z^T using CLIP's text encoder, as described in Section 3.1. We have:

$$z^I = f_I(I) \in \mathbb{R}^{512}, \qquad z^T = f_T\left(\text{prompt}(t_{\text{obj}}, t_{\text{ctx}}, t^{(1)}_{\text{smell}}, \ldots, t^{(K)}_{\text{smell}})\right) \in \mathbb{R}^{512}.$$

While the text encoder remains frozen to preserve semantic structure, the image encoder is fine-tuned to adapt its features towards olfactory-relevant cues.

Multimodal Alignment. We train the smell encoder by aligning the projected smell representations with their corresponding visual and textual embeddings using symmetric InfoNCE losses [42, 45]. Let $\mathrm{sim}(\cdot,\cdot)$ denote cosine similarity. For a batch of N samples we compute

$$\mathcal{L}_{IS} = -\frac{1}{2N}\sum_{i=1}^{N}\left(\log\frac{\exp(\mathrm{sim}(z^S_{I,i}, z^I_i)/\tau)}{\sum_{j=1}^{N}\exp(\mathrm{sim}(z^S_{I,i}, z^I_j)/\tau)} + \log\frac{\exp(\mathrm{sim}(z^S_{I,i}, z^I_i)/\tau)}{\sum_{j=1}^{N}\exp(\mathrm{sim}(z^S_{I,j}, z^I_i)/\tau)}\right)$$

where τ is a learnable temperature parameter. $\mathcal{L}_{ST}$ is computed analogously using (z^S_T, z^T). The overall representation learning objective is

$$\mathcal{L}_{\text{total}} = \lambda_{IS}\mathcal{L}_{IS} + \lambda_{ST}\mathcal{L}_{ST} \quad . \tag{1}$$

This stage produces an olfactory embedding that is jointly structured by the smell measurements, the visual scene, and the augmented language descriptors generated by the VLM.

3.3 Latent Smell Disentanglement

The representation z^S learned in the previous stage captures the overall olfactory signal present in the measurement. However, smell recordings often contain a mixture of sources, including odors emitted from the target object as well as background environmental factors. To disentangle these contributions, we introduce a latent decomposition of the olfactory representation into object-specific and background-related components. This step is illustrated in Figure 3b.

Latent Decomposition. Our goal is to decompose the learned embedding into two components:

$$z^S_{\text{obj}} = \phi_{\text{obj}}(z^S_T) \qquad \text{and} \qquad z^S_{\text{ctx}} = \phi_{\text{ctx}}(z^S_T) \quad ,$$

where z^S_{obj} and z^S_{ctx} *exclusively* represent object-related odor factors and the remaining contextual environmental contributions, respectively.

Semantic Alignment. The decomposition is enabled by the practical compositionality of language descriptors, i.e. we have separate text encodings for an 'object' and the 'context' of the scene it is located in. During training, each decoded smell component is aligned with its corresponding textual descriptor obtained in Section 3.1 to provide explicit supervision: the object component z^S_{obj} is aligned with the object descriptor t_{obj}, while the contextual component z^S_{ctx} is aligned with the context descriptor t_{ctx}.

Denoting z^T_{obj} and z^T_{ctx} the corresponding CLIP text embeddings. We enforce this decomposition through contrastive losses. Let $\mathrm{sim}(\cdot,\cdot)$ denote cosine similarity. For a batch of N samples, we compute:

$$\mathcal{L}_{obj} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(z^S_{obj,i}, z^T_{obj,i})/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(z^S_{obj,i}, z^T_{obj,j})/\tau)}$$

$$\mathcal{L}_{ctx} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(z^S_{ctx,i}, z^T_{ctx,i})/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(z^S_{ctx,i}, z^T_{ctx,j})/\tau)}.$$

Reconstruction Constraint. To ensure that the decomposition preserves the information contained in the original smell signal without representation collapse, we also implement a reconstruction objective. A decoder $d(\cdot)$ predicts the smell signal from the concatenated latents:

$$\mathcal{L}_{\text{rec}} = \|\mathbf{X} - d([z^S_{\text{obj}}, z^S_{\text{ctx}}])\|^2 \quad .$$

The disentanglement stage is therefore a weighted combination of these losses:

$$\mathcal{L}_{\text{dis}} = \mathcal{L}_{\text{obj}} + \mathcal{L}_{\text{ctx}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} \quad . \tag{2}$$

λ_{rec} is a weighting factor for the reconstruction loss, which acts as a regularizer to prevent embedding collapse. This decomposition encourages SCENT to separate object-specific odor signals from contextual environmental contributions while remaining consistent with the underlying sensor measurements.

4 Experiments

4.1 Implementation details and experiment settings

Dataset. We evaluate SCENT on the **New York Smells (NYS)** dataset [44], the largest multimodal benchmark for in-the-wild olfactory perception. NYS contains 7,000 paired image-smell samples across 3,500 distinct object categories. We use 5,996 training and 936 validation samples; all baselines and ablations share this split.

Implementation Details. Each olfactory sample comprises a 32-channel resistance signal from an e-nose array, consisting of a baseline recording ($\mathbf{B} \in \mathbb{R}^{C\times T}$) of ambient air and a sample recording ($\mathbf{S} \in \mathbb{R}^{C\times T}$) taken near the object, where $C = 32$ is the number of sensor channels and T denotes the temporal dimension (typically 14 timesteps). Following [44], we use a unified olfactory input $\mathbf{X}$ formed by the temporal concatenation of the two signals: $\mathbf{X} = [\mathbf{B}; \mathbf{S}] \in \mathbb{R}^{C\times 2T}$. Our olfactory encoder f_S is a Transformer with 6 layers and 8 attention heads. The input signal $\mathbf{X}$ is projected to a 448 dimensional embedding and augmented with sinusoidal positional encodings, the resulting olfactory representation z^S is then mapped to the 512-dimensional CLIP visual and textual spaces using two MLP projection heads. For semantic augmentation, we use Qwen3VL-30B [65] to generate descriptors $\{t_{\text{obj}}, t_{\text{ctx}}, t_{\text{smell}}\}$. We employ CLIP ViT-B/16 [45] as our image-language encoder backbone. During training, the text encoder remains

frozen, while the vision encoder is trainable, unless stated otherwise in specific ablation studies. Additional architectural and training details are provided in the supplementary material.

Evaluation Metrics and Retrieval Tasks. We evaluate SCENT across two primary categories, single-modality and joint-modality retrieval, to assess the alignment of olfactory features with individual and combined semantic spaces. We report standard retrieval metrics: Recall@k (%) (R@k for $k \in \{1, 5, 10, 20\}$).
Single-modality Retrieval (S2I, S2T): These tasks evaluate the model's ability to align olfactory signals with a single target modality. In Smell-to-Image (S2I) and Smell-to-Text (S2T), a query smell signal is used to rank and retrieve candidates from a set of images or textual descriptors, respectively, based on cross-modal similarity.
Joint-modality Retrieval (S2IT): This task evaluates the model's performance in a fully multimodal setting. The query is a singular olfactory signal, and the search database consists of (Image, Text) pairs. For each candidate pair in the gallery, the model must compute a joint similarity score that accounts for both the visual and textual components. We compute this score via latent-level fusion of the dual olfactory projections; full details and a comparison with an alternative fusion strategy are provided in the supplementary material.

Baselines. As the current state-of-the-art on the NYS dataset [44] is a vision-only model, it serves as our primary comparison for the S2I task. However, since this baseline lacks a native textual projection, it cannot be directly applied to tasks involving language. Moreover, the original work does not provide publicly available code or pretrained weights, requiring us to reproduce the model based on the descriptions in the paper. To establish a rigorous baseline for text-based tasks, we reproduce the NYS model and implement NYS (adapted) for multimodal retrieval tasks, which utilizes a two-stage "Image Bridge" (IB) protocol:

- **S2T Retrieval:** The model first retrieves the top-1 most visually similar image to the query smell using its olfactory-visual head. This "bridge" image is then encoded via a frozen CLIP ViT-B/16 [45] model and used to rank textual candidates in the CLIP latent space.
- **Joint Retrieval (S2IT):** For S2IT, a composite score is formed by summing the smell-image similarity (from the NYS model) with the pre-existing image-text alignment score (from CLIP) for each candidate pair.

4.2 The Necessity of Language Guidance

We first motivate SCENT by evaluating the impact of different supervision levels while keeping visual and textual backbones frozen (Table 1). Aligning smell solely with images (S2I) achieves a baseline R@5 of 12.1. When aligning strictly with text (S2T), we observe that performance is highly dependent on the semantic granularity of the descriptors. Using only isolated object labels (**O**) results in poor alignment (8.8 R@5), as these labels fail to capture the environmental

Table 1: Impact of semantics on olfactory alignment. We evaluate the retrieval performance of our olfactory encoder when aligned with frozen CLIP visual and textual spaces. Results indicate that simple object labels (**O**) provide insufficient supervision, whereas incorporating context (**Ctx**), and inferred ambient smells (**I**) yields the better cross-modal alignment across both single and joint retrieval tasks. Best results in **bold**.

Align. Strategy	**Textual Info**			**Smell Retrieval**		
	O	**Ctx**	**S**	R@5	R@10	R@20
Image (S2I)				12.1	18.3	29.2
Text (S2T)	✓			8.8 (-3.3)	14.4 (-3.9)	24.9 (-4.3)
	✓	✓		10.8 (-1.3)	18.3 (0.0)	28.1 (-1.1)
	✓	✓	✓	12.4 (+0.3)	19.2 (+0.9)	29.0 (-0.2)
Image & Text (S2IT)	✓	✓	✓	**14.3** (+2.2)	**22.5** (+4.2)	**35.1** (+5.9)

factors present in the sensor data. However, as we incrementally add contextual descriptors (**Ctx**) and VLM-inferred ambient smells (**S**), performance improves significantly, eventually surpassing the image-only baseline. The most robust representations are formed through the joint **Image&Text** strategy. By aligning the smell encoder with both modalities simultaneously using our semantic-rich template (**O**+**Ctx**+**S**), we achieve 14.3 R@5. This confirms that language guidance offers a critical semantic bridge that visual features alone cannot provide.

4.3 Comparison to State-of-the-Art

In Table 2, we compare SCENT against the state-of-the-art NYS benchmark.

S2I Retrieval. We observe that language guidance improves performance even on purely visual tasks. In the S2I setting, SCENT outperforms the reproduced NYS baseline (R@5 from 20.0 to 22.1) even only with object labels **O**. This confirms that the semantic structure of language helps the olfactory encoder extract more discriminative features than vision alone. As contextual (**Ctx**) and inferred smell (**S**) descriptors are added, S2I performance peaks at 23.0 R@5, indicating that a more complete semantic description of the scene complements the visual signal and improves the olfactory representation.

S2T and S2IT Retrieval. On tasks involving language, the advantage of our proposed architecture is even more apparent. The NYS (adapted) baseline struggles with the semantic gap, as its "Image Bridge" is limited by what is explicitly visible. In contrast, SCENT leverages its native textual projection head to achieve 11.9 R@5 on S2T retrieval, a 47% relative improvement over the best adapted baseline. Furthermore, our joint-modality retrieval (S2IT) results show that SCENT effectively fuses visual and textual cues, outperforming the baseline by over 7% in R@5.

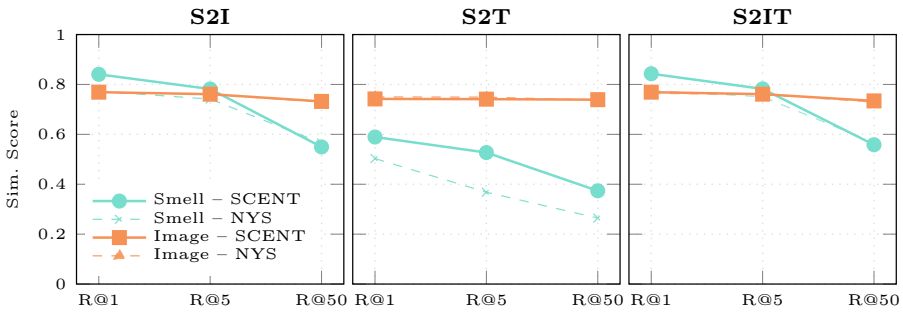


Fig. 4: Mean cosine similarity between the ground-truth and top-k retrieved items, in frozen CLIP image space (orange) and learned olfactory space (teal), for $k \in \{1, 5, 50\}$. Solid lines: SCENT; dashed: NYS.

Visual vs. Olfactory Discriminativity. Figure 4 plots the mean cosine similarity between the ground-truth item and the top-k retrieved items, measured in frozen CLIP image space (orange) and in the learned olfactory space (teal), for $k \in \{1, 5, 50\}$. Image features are weakly discriminative: scores remain nearly flat across retrieval ranks, failing to separate good retrievals (R@1) from bad ones (R@50). In contrast, olfactory features drop sharply with retrieval rank; SCENT's smell curve (solid) decays faster than NYS's (dashed), confirming both that smell is the necessary modality for this task and that SCENT learns a more discriminative representation. Notably, NYS achieves similar S2T visual similarity to SCENT yet 32% lower R@5, showing that visual proximity does not imply correct olfactory match.

Qualitative Analysis. We visualize the S2IT retrieval performance in Figure 5. Each row corresponds to one example and the columns illustrate the top-5 images from the retrieved (Image,Text) pairs. Despite the inherent difficulty of the olfactory-to-visual mapping, SCENT consistently ranks the correct semantic category within the top-5 results. Notably, the model handles fine-grained differences (e.g., distinguishing between different types of foliage or tree barks) and identifies objects regardless of their visual state, such as the peeled banana or the sliced pizza, by leveraging the underlying olfactory cue.

4.4 Ablation Studies

Impact of Semantic Granularity. We first evaluate how the complexity of the textual descriptors impacts olfactory alignment. As shown in Table 2, we observe a consistent performance gain across all tasks as the supervision transitions from simple object labels (**O**) to scene-aware descriptions that include environmental context (**Ctx**) and inferred ambient smells (**S**). In the **S2I** task, expanding the semantic scope beyond basic labels improves R@5 from 22.1 to 23.0, suggesting that language context complements the missing information from the visual

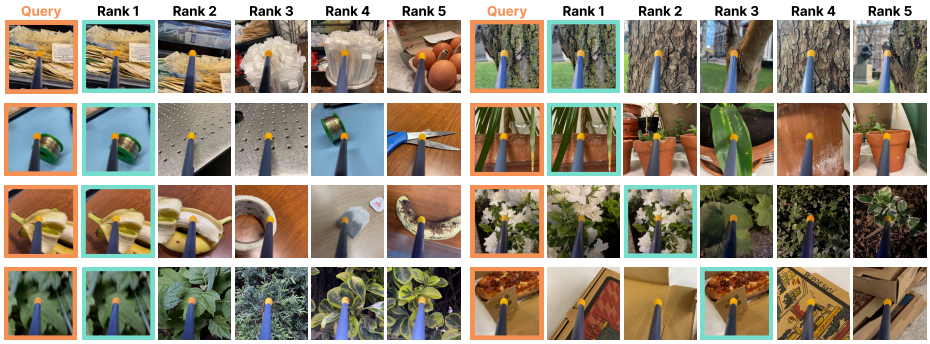


Fig. 5: Qualitative results for S2IT retrieval. For a given *olfactory* query, we show the ground-truth image in the **Query** column and the top-5 images from the retrieved (Image, Text) pairs. The **Correct** retrieval is highlighted. SCENT demonstrates robust cross-modal alignment across diverse categories, successfully retrieving the correct match even when the visual appearance varies significantly from the query (e.g., the pizza box in Rank 1 vs. the open pizza in query).

modality during olfactory representation training. This trend is even more evident in language-centric tasks: for **S2T**, moving from **O** to the full **O** + **Ctx** + **S** template yields an improvement from 8.0 to 11.9 R@5. Notably, the inclusion of VLM-inferred smells provides the best performance, validating our hypothesis that while visual features may miss the olfactory essence of a scene, high-level language inferring can recover these unobservable cues to better align the olfactory latent space.

Image Bridge Ablation. To isolate the benefit of our three-stream architecture from the VLM-generated language guidance, we apply the Image Bridge protocol to SCENT: we disable the ST head and route S2T/S2IT retrieval via the IS head and a frozen CLIP proxy, mirroring the NYS† (adapted) baseline (Table 3). S2T and S2IT degrade, confirming that the gains stem from the three-stream design and not solely from VLM-enriched training data. Notably, SCENT with IB still outperforms NYS† (adapted) with IB, showing that co-training with the ST loss also improves the IS head.

VLM Sensitivity Analysis We evaluate how the choice of Vision-Language Model affects retrieval performance, keeping all other components of SCENT fixed. Results are shown in Table 5. Performance scales with VLM reasoning quality rather than parameter count alone: Qwen3-VL-30B outperforms the larger Qwen2.5-VL-72B across all tasks, consistent with the multi-hop reasoning required to infer plausible olfactory cues from visual context. Smaller models (Qwen3-VL-8B, Gemma 4 E4B) produce weaker annotations, yielding lower retrieval performance, particularly on language-centric tasks (S2T). These results confirm that annotation richness, and therefore downstream alignment quality, is driven primarily by the model's reasoning capability rather than its scale alone.

Table 2: State-of-the-art comparison on the NYS dataset. We compare SCENT against the vision-only NYS baseline [44] across single-modality (S2I, S2T) and joint-modality (S2IT) tasks. To enable a comparison on language-based tasks, we implement NYS (adapted) using an "Image Bridge" protocol. Our method consistently outperforms the adapted baseline, demonstrating that native three-stream alignment with inferred semantic cues captures nuanced olfactory information that is unobservable through visual features alone.

Method	Textual Info			S2I				S2T				S2IT			
	O	Ctx	S	R@1	R@5	R@10	R@20	R@1	R@5	R@10	R@20	R@1	R@5	R@10	R@20
NYS* [44]				—	16.5	29.6	43.1			—				—	
NYS†				5.4	20.0	29.9	42.0			—				—	
NYS† (adapted)	✓					—		1.7	6.2	10.4	14.4	3.9	15.6	25.8	39.1
	✓	✓				—		2.5	8.7	12.7	18.3	3.9	15.9	25.2	38.7
	✓	✓	✓			—		2.6	8.1	12.9	19.2	4.9	18.5	27.5	40.7
Ours	✓			5.6	22.1	32.7	42.4	1.6	8.0	13.8	21.5	5.9	22.0	32.4	42.6
	✓	✓		5.9	21.8	32.4	**45.2**	2.8	10.7	17.3	28.1	6.1	21.3	32.5	**45.7**
	✓	✓	✓	**6.0**	**23.0**	**33.5**	43.6	**3.3**	**11.9**	**19.8**	**29.2**	**6.8**	**23.3**	**32.7**	42.5

* Results reported in the original NYS paper [44].
† Results reproduced.

4.5 Annotation quality.

To verify that VLM-generated descriptors infer real olfactory context rather than plausible hallucinations, we use a second VLM, the "judge". We provide a held-out second view of the same scene [1], and the judge scores each annotation on two criteria: **plausibility** (is the annotation contextually reasonable given both views?), reaching 98.6%; and **View-2 confirmation** (does the annotation specifically predict content hidden in View 1 but visible in View 2?), reaching 32.2%. This confirms that our descriptors capture real concealed scene content rather than generic guesses, as illustrated in Figure 6.

4.6 Latent Smell Disentanglement

We evaluate the ability of SCENT to decompose a complex olfactory signal into its constituent parts: the target object and the environmental context.

Evaluation Protocol: Recombination Retrieval. To evaluate whether the latent decomposition ϕ_{obj} and ϕ_{ctx} effectively isolates object-specific and environmental olfactory components, we propose a **zero-shot recombination retrieval** task. We identify (Object, Context) pairs in the validation set that never appear

[1] The NYS dataset provides two different camera views per recording, captured simultaneously from different angles of the same scene (Figure 1). View 2 is never used during the annotation process or in training. It is withheld exclusively for this validation experiment.

Table 3: Image Bridge (IB) ablation. Replacing SCENT's native ST head with an Image Bridge proxy confirms that performance gains stem from the three-stream architecture.

Method	IB	S2T		S2IT	
		R@5	R@20	R@5	R@20
NYS† (adapted)	✓	8.1	19.2	18.5	40.7
SCENT	✓	9.1	18.8	21.4	43.8
SCENT (ours)	✗	**11.9**	**29.2**	**23.3**	**42.5**

Table 4: Recombination retrieval. Decoded Synthesis (ours) outperforms **Raw Recombination** on zero-shot (Object, Context) pairings unseen during training.

Query Type	R@1	R@5	R@10	R@20
Raw Recombination	**3.9**	5.9	7.8	9.8
Decoded Synthesis (ours)	2.0	**9.8**	**11.8**	**13.7**

Table 5: VLM sensitivity analysis. All variants are SCENT trained with **O**+**Ctx**+**S** descriptors; only the VLM used for annotation differs. MMLU score is reported as a proxy for general reasoning capability.

VLM	MMLU	S2I		S2T		S2IT	
		R@5	R@20	R@5	R@20	R@5	R@20
Gemma 4 E4B [24]	69.4	19.6	43.3	7.7	19.8	20.2	44.4
Qwen2.5-VL-72B [3]	71.2	20.0	42.1	8.1	20.5	19.7	41.9
Qwen3-VL-8B [65]	71.6	19.1	39.9	6.5	22.0	20.3	40.0
Qwen3-VL-30B [65] (ours)	**77.8**	**23.0**	**43.6**	**11.9**	**29.2**	**23.3**	**42.5**

together in the training data, although the individual objects and contexts are present separately across different training samples.

For each such pair, we construct a synthetic query by pairing an object factor z^S_{obj} from one training sample (e.g., "coffee") with a context factor z^S_{ctx} from a distinct training sample with the target background (e.g., "outdoor park"). We then utilize the decoder $g(\cdot)$ to synthesize a novel smell fingerprint $\hat{\mathbf{X}} = g([z^S_{\text{obj}}; z^S_{\text{ctx}}])$. The task is to retrieve the ground-truth validation samples where this specific combination occurred.

We compare our **Decoded Synthesis** against a **Raw Recombination** baseline, which creates a query by manually concatenating the raw sensor baseline (**B**) and sample (**S**) signals from two different training exemplars. This protocol tests the model's ability to perform compositional generalization.

Quantitative Analysis. As shown in Table 4, we evaluate the retrieval performance of these synthetic queries within our latent space, by passing the query smell fingerprint $\mathbf{X}$ through our smell encoder. Our Decoded Synthesis outperforms the Raw Recombination baseline. While the raw baseline plateaus at 9.8 R@20, our synthesized fingerprints achieve 13.7 R@20.

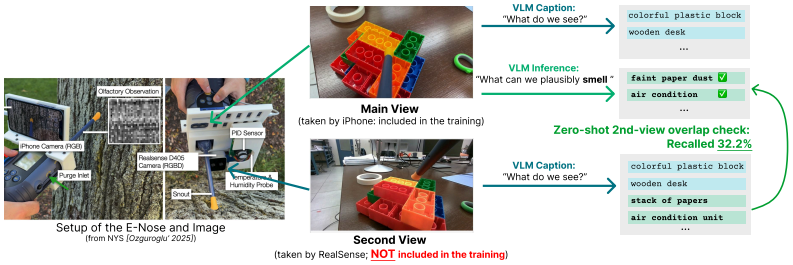


Fig. 6: Held-out view validation setup. A VLM infers plausible smells from View 1; a second VLM judge uses a held-out View 2 to verify the inferences, distinguishing grounded predictions from hallucinations.

5 Limitations and Conclusion

Despite the performance gains of SCENT, several challenges remain. The reliance on e-nose sensors introduces inherent stochasticity, as these hardware devices are prone to temporal drift and cross-sensitivity to environmental factors such as humidity and temperature. Furthermore, while the NYS dataset is a significant milestone, a substantial gap remains between olfactory benchmarks and the web-scale datasets used to train vision-language backbones.

In this paper, we addressed the fundamental challenge of aligning high-dimensional electronic-nose signals with visual data, recognizing that images often fail to capture the pervasive environmental context of olfaction. We introduced SCENT, a multimodal framework that employs language guidance as a semantic bridge, leveraging the extensive world knowledge embedded in VLMs to infer plausible ambient olfactory cues. Through a language-guided latent decomposition, our model effectively disentangles object-specific odors from broader environmental factors. Extensive experiments on the New York Smells (NYS) dataset demonstrate that this semantic grounding enhances cross-modal retrieval performance across both single and joint modality tasks. Moroever, our language-guidance mechanism achieves decompositionality of olfactory information, facilitating generalization to out-of-distribution (OOD) samples.

Acknowledgements

This work is supported by Hi! Paris and the ANR/France 2030 program (ANR-23-IACL-0005), a Hi! Paris grant and fellowship, a CIEDS grant, and a Google DeepMind academic gift. Computing resources were provided by GENCI through access to the IDRIS High-Performance Computing facilities under allocations 2026-AD011014300R3 and 2025-AD011015893R1, and by Google Gemini. We sincerely thank Mathieu Aubry and the anonymous reviewers for their insightful discussions that contributed to this work. We are also grateful to Julie Mordacq and Robin Courant for their meticulous proofreading.

References

1. Ali, M.M., Hashim, N., Abd Aziz, S., Lasekan, O.: Principles and recent advances in electronic nose for quality inspection of agricultural and food products. Trends in Food Science & Technology (2020)
2. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: NeurIPS (2016)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv: 2502.13923 (2025)
4. Bender, T., Sørensen, S., Kashani, A., Eldjarn Hjorleifsson, K., Hyldig, G., Hauberg, S., Belongie, S., Warburg, F.: Learning to taste: A multimodal wine dataset. In: NeurIPS (2023)
5. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al.: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164 (2024)
6. Brattoli, M., De Gennaro, G., De Pinto, V., Loiotile, A.D., Lovascio, S., Penza, M.: Odour detection methods: Olfactometry and chemical sensors. Sensors (2011)
7. Chen, H., Huo, D., Zhang, J.: Gas recognition in e-nose system: A review. IEEE transactions on biomedical circuits and systems (2022)
8. Chen, S., Li, H., Wang, Q., Zhao, Z., Sun, M.T., Zhu, X., Liu, J.: VAST: A vision-audio-subtitle-text omni-modality foundation model and dataset. In: NeurIPS (2023)
9. Cicchetti, G., Grassucci, E., Comminiello, D.: A triangle enables multimodal alignment beyond cosine similarity. In: NeurIPS (2025)
10. Cicchetti, G., Grassucci, E., Sigillo, L., Comminiello, D.: Gramian multimodal representation learning and alignment. In: ICLR (2025)
11. De Sa, V.: Learning classification with unlabeled data. In: NeurIPS (1993)
12. Debnath, T., Nakamoto, T.: Predicting human odor perception represented by continuous values from mass spectra of essential oils resembling chemical mixtures. PloS one (2020)
13. Dou, Y., Yang, F., Liu, Y., Loquercio, A., Owens, A.: Tactile-augmented radiance fields. In: CVPR (2024)
14. Dravnieks, A.: Atlas of Odor Character Profiles. ASTM Special Technical Publication (1985)
15. Dufumier, B., Castillo Navarro, J., Tuia, D., Thiran, J.P.: What to align in multimodal contrastive learning? In: ICLR (2025)
16. Elizalde, B., Deshmukh, S., Al Ismail, M., Wang, H.: CLAP: learning audio concepts from natural language supervision. In: ICASSP (2023)
17. Erlangga, F., Wijaya, D.R., Wikusna, W.: Electronic nose dataset for classifying rice quality using neural network. In: International Conference on Information and Communication Technology (ICoICT) (2021)
18. Fang, X., Mao, K., Duan, H., Zhao, X., Li, Y., Lin, D., Chen, K.: Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. NeurIPS (2024)
19. 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)
20. Fundurulic, A., Faria, J.M., Inácio, M.L.: Advances in electronic nose sensors for plant disease and pest detection. Engineering Proceedings (2023)

21. Furizal, F., Ma'arif, A., Firdaus, A.A., Rahmaniar, W.: Future potential of e-nose technology: A review. International Journal of Robotics and Control Systems (2023)
22. Ghazaly, C., Biletska, K., Thevenot, E.A., Devillier, P., Naline, E., Grassin-Delyle, S., Scorsone, E.: Assessment of an e-nose performance for the detection of covid-19 specific biomarkers. Journal of Breath Research (2023)
23. 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)
24. Google DeepMind: Gemma model documentation. `https://ai.google.dev/gemma/docs/core` (2026), accessed: 2026-06-30
25. Han, J., Gong, K., Zhang, Y., Wang, J., Zhang, K., Lin, D., Qiao, Y., Gao, P., Yue, X.: Onellm: One framework to align all modalities with language. In: CVPR (2024)
26. Huh, M., Cheung, B., Wang, T., Isola, P.: The platonic representation hypothesis. arXiv preprint arXiv:2405.07987 (2024)
27. 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)
28. Keller, A., Gerkin, R.C., Guan, Y., Dhurandhar, A., Turu, G., Szalai, B., Mainland, J.D., Ihara, Y., Yu, C.W., Wolfinger, R., et al.: Predicting human olfactory perception from chemical features of odor molecules. Science (2017)
29. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E.P., Sanketi, P.R., Vuong, Q., et al.: Openvla: An open-source vision-language-action model. In: 8th Annual Conference on Robot Learning (2025)
30. Kontaris, I., East, B.S., Wilson, D.A.: Behavioral and neurobiological convergence of odor, mood and emotion: A review. Frontiers in Behavioral Neuroscience (2020)
31. Koutoupis, S., Zervou, M.A., Kontras, K., De Vos, M., Tsakalides, P., Tsagkatakis, G.: The more, the merrier: Contrastive fusion for higher-order multimodal alignment. In: CVPR (2026)
32. 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)
33. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. IEEE Transactions on Machine Learning Research (2025)
34. 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)
35. Liu, J., Chen, S., He, X., Guo, L., Zhu, X., Wang, W., Tang, J.: Valor: Vision-audio-language omni-perception pretraining model and dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
36. Liu, Y., Li, S., Wu, Y., Chen, C.W., Shan, Y., Qie, X.: Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In: CVPR (2022)
37. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. Neurocomputing (2022)
38. Makri, E., Nakis, N., Sisson, L., Minsky, G., Tassiulas, L., Satarifard, V., Christakis, N.A.: Benchmark for assessing olfactory perception of large language models. arXiv preprint arXiv:2604.00002 (2026)

39. Moein, S.T., Hashemian, S.M., Mansourafshar, B., Khorram-Tousi, A., Tabarsi, P., Doty, R.L.: Smell dysfunction: a biomarker for covid-19. In: International forum of allergy & rhinology (2020)
40. Mueller, P., Salminen, K., Nieminen, V., Kontunen, A., Karjalainen, M., Isokoski, P., Rantala, J., Savia, M., Väliaho, J., Kallio, P., et al.: Scent classification by k nearest neighbors using ion-mobility spectrometry measurements. Expert systems with applications (2019)
41. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y., et al.: Multimodal deep learning. In: ICML (2011)
42. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
43. Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E.H., Freeman, W.T.: Visually indicated sounds. In: CVPR (2016)
44. 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)
45. 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)
46. Ravia, A., Snitz, K., Honigstein, D., Finkel, M., Zirler, R., Perl, O., Secundo, L., Laudamiel, C., Harel, D., Sobel, N.: A measure of smell enables the creation of olfactory metamers. Nature (2020)
47. Rodríguez, J., Durán, C., Reyes, A.: Electronic nose for quality control of colombian coffee through the detection of defects in "cup tests". Sensors (2009)
48. Ruan, L., Hu, A., Song, Y., Zhang, L., Zheng, S., Jin, Q.: Accommodating audio modality in clip for multimodal processing. In: AAAI (2023)
49. 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)
50. Saporta, A., Puli, A.M., Goldstein, M., Ranganath, R.: Contrasting with symile: Simple model-agnostic representation learning for unlimited modalities. In: NeurIPS (2024)
51. Snitz, K., Yablonka, A., Weiss, T., Frumin, I., Khan, R.M., Sobel, N.: Predicting odor perceptual similarity from odor structure. PLoS computational biology (2013)
52. Stevenson, R.J.: An initial evaluation of the functions of human olfaction. Chemical senses (2010)
53. Tan, J., Xu, J.: Applications of electronic nose (e-nose) and electronic tongue (e-tongue) in food quality-related properties determination: A review. Artificial Intelligence in Agriculture (2020)
54. Torres-Tello, J., Guaman, A.V., Ko, S.B.: Improving the detection of explosives in a mox chemical sensors array with lstm networks. IEEE Sensors Journal (2020)
55. Vergara, A., Vembu, S., Ayhan, T., Ryan, M.A., Homer, M.L., Huerta, R.: Chemical gas sensor drift compensation using classifier ensembles. Sensors and Actuators B: Chemical (2012)
56. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: CVPR (2015)
57. Wachowiak, M., Dewan, A., Bozza, T., O'Connell, T.F., Hong, E.J.: Recalibrating olfactory neuroscience to the range of naturally occurring odor concentrations. Journal of Neuroscience (2025)
58. Wang, W., Nie, A., Zhou, W., Kai, Y., Hu, C.: Teaching physical awareness to llms through sounds. In: ICML (2025)

59. Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., Shi, Y., et al.: Internvideo2: Scaling foundation models for multimodal video understanding. In: ECCV (2024)
60. Wijaya, D.R., Sarno, R., Zulaika, E.: Dwtlstm for electronic nose signal processing in beef quality monitoring. Sensors and Actuators B: Chemical (2021)
61. Wilson, A.D.: Review of electronic-nose technologies and algorithms to detect hazardous chemicals in the environment. Procedia Technology (2012)
62. Xiao, J., Yao, A., Li, Y., Chua, T.S.: Can i trust your answer? visually grounded video question answering. In: CVPR (2024)
63. Xu, H., Ye, Q., Yan, M., Shi, Y., Ye, J., Xu, Y., Li, C., Bi, B., Qian, Q., Wang, W., et al.: mplug-2: A modularized multi-modal foundation model across text, image and video. In: ICML (2023)
64. Xue, L., Yu, N., Zhang, S., Panagopoulou, A., Li, J., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J.C., et al.: Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In: CVPR (2024)
65. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
66. 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)
67. Yu, J., Wang, X., Tu, S., Cao, S., Zhang-Li, D., Lv, X., Peng, H., Yao, Z., Zhang, X., Li, H., et al.: Kola: Carefully benchmarking world knowledge of large language models. In: ICLR (2024)
68. Yuan, W., Wang, S., Dong, S., Adelson, E.: Connecting look and feel: Associating the visual and tactile properties of physical materials. In: CVPR (2017)
69. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., et al.: Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106 (2025)
70. Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., Li, H.: Pointclip: Point cloud understanding by clip. In: CVPR (2022)
71. Zhao, Y., Misra, I., Krähenbühl, P., Girdhar, R.: Learning video representations from large language models. In: CVPR (2023)
72. Zhou, B., Li, L., Wang, Y., Liu, H., Yao, Y., Wang, W.: Unialign: Scaling multimodal alignment within one unified model. In: CVPR (2025)
73. Zhu, B., Lin, B., Ning, M., Yan, Y., Cui, J., HongFa, W., Pang, Y., Jiang, W., Zhang, J., Li, Z., et al.: Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In: ICLR (2024)