What Images Cannot Say: Language-Guided Olfactory Representation Learning

Eleftherios Tsonis[1], Xi Wang[1], and Vicky Kalogeiton[1]

LIX, École Polytechnique, IP Paris, CNRS

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 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.

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..*

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 [29, 30, 46], answer questions about images [14, 52, 57], and reason about complex environments through language [4, 26]. From this perspective, one might conclude that images already capture nearly everything that can be inferred about the world from visual observation.

Yet despite this progress, current models primarily capture *what a scene looks like*, while remaining largely unaware of what it might *feel like to be there*.

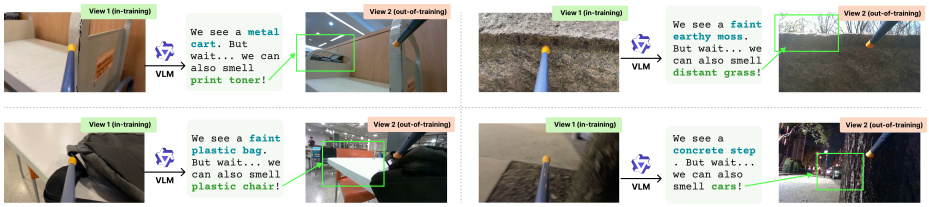


Fig. 1: Not everything we can smell is visible. Given only View 1 (in-sample) during training, the true olfactory context may lie outside the field of view, in View 2 (out-of-sample). A VLM can bridge this gap by inferring plausible smells from semantic context alone. We use these language-derived signals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

In real environments, perception extends beyond vision. A photograph of a busy street may suggest the smell of exhaust; a metro entrance may imply the metallic scent of ventilation air. Such environmental cues are rarely visible directly, but humans routinely infer them from visual context.

We argue that understanding real environments requires going beyond describing what we see to reasoning about the physical signals surrounding a scene. Instead, current multimodal systems primarily operate on representations derived from vision and language, while largely ignoring the **physical signals** that characterize real environments, such as haptics [20, 28], acoustics [1, 17] etc. Among these signals, olfaction is particularly important. Smell provides rich information about environmental conditions [5, 19, 45], including pollution levels, food quality, and potential hazards.

Advances in electronic noses (e-noses) [51] enable the capture of high-dimensional sensor measurements [13, 15, 38] that characterize the chemical composition of air. However, interpreting these signals remains challenging. Olfactory measurements typically represent mixtures of multiple odor sources and depend strongly on environmental context, making them difficult to interpret in isolation.

Our key observation is that language can provide a semantic bridge between visual scenes and olfactory signals. While smell itself is not visible, many olfactory cues are strongly correlated with contextual elements present in a scene. For example, a bakery storefront implies the aroma of fresh bread, while indoor plants may indicate humidity or soil-related smells. Modern VLMs encode extensive **world knowledge** about objects and environments, enabling them to generate semantic descriptions that expose such contextual cues. These descriptions can therefore provide useful guidance for interpreting olfactory measurements that would otherwise be difficult to understand in isolation (Figure 1).

Building on this insight, we propose **SCENT**: **S**emantic **C**ontext-aware **e-N**ose **T**ransformer, a framework that integrates electronic-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 descriptors are then encoded using a pretrained VLM and used as semantic anchors for representa-

tion learning. We train a smell encoder that maps electronic-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 [38] and demonstrate that integrating textual scene descriptors 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 as a semantic bridge** between visual scenes and olfactory signals, enabling the interpretation of smell measurements through contextual world knowledge.
- We propose a **multimodal framework for learning olfactory representations** that aligns electronic-nose signals with visual and textual embeddings.
- We introduce a **language-guided decomposition of olfactory representations** that separates object-related odors from contextual environmental contributions.

2 Related Work

Cross-modal and multimodal supervision. Using one modality to supervise another has a long history: from early self-supervision with hearing for vision [9] and deep multimodal models [35], to audio-visual [2,37] and visual-tactile [11,55,56] alignment from co-occurring sensor data. Language has become a central anchor for vision and audio: CLIP [39] and ALIGN [24] align image and text; CLAP [12] aligns audio and text; extensions include video-text (e.g., CLIP4Clip [33]) and point cloud–text [58]. For three modalities, current systems choose an anchor (often image, audio, or text) and align the others to it via pairwise cosine similarity: ImageBind [22] uses image as the bridge; LanguageBind [60] uses text; CLIP4VLA [41] incorporates audio with video and text. Large tri-modal models (VALOR [31], VAST [6], mPLUG-2 [53], InternVideo2 [50]) follow similar pairwise schemes. At inference, methods often fuse two modalities and compare to the third, or use extra fusion layers [6, 32, 53]. Recent work has proposed methods for n-modal alignment (e.g., GRAM [8], Symile [43], TRIANGLE [7], OneLLM [23]).

Language generation as semantic supervision. Generated language can act as scalable latent supervision to ground embeddings in high-level scene seman-

tics. Image caption generation provides a natural interface for semantic supervision [48]. Recent work extends this idea by generating language to supervise hard-to-label modalities: ULIP-2 [54] uses multimodal models to produce holistic 3D descriptions for scalable tri-modal pretraining, while LAVILA [59] repurposes Large Language Models as narrators to densely annotate video and improve contrastive video–text learning. Furthermore, taste modelling in WineSensed/FEAST [3] shows that text and images can be aligned to human sensory spaces. 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, often requiring specialized or laboratory-grade sensing [10, 16, 34], with applications in scent design [42], disease detection [18, 21], and security [47]. Research at the molecular level has used psychophysical data and graph-based models to predict perceptual attributes [25] and to define a principal odor map (POM) [27], while mixture and similarity studies remain limited [40, 44]. Recent work in neuroscience stresses the need for data in natural concentration ranges [49]. Raw e-nose signals are high-dimensional and noisy, making data-driven representation learning attractive.

Moving beyond lab-based limitations, the New York Smells (NYS) dataset [38] is a large-scale multimodal benchmark designed for "in-the-wild" smell perception. NYS comprises 7,000 image and olfactory pairs. The samples are collected across diverse urban environments, ranging from indoor libraries to outdoor urban parks, capturing a wide array of ambient scents and environmental noise. To our knowledge, no existing work studies multimodal learning that combines *olfactory sensor signals* with *text* (or vision and text together), which is the setting we address.

3 Method

We study the problem of learning semantically aligned representations of smell from electronic-nose measurements. Given an image I of a scene and a corresponding smell signal $\mathbf{X}$, also referred to as a smell fingerprint, recorded by an electronic nose, our goal is to learn an olfactory encoder f_S that maps the smell signal to a representation:

$$z^S = f_S(\mathbf{X}) \quad , \tag{1}$$

that captures the complete semantic structure of the underlying olfactory environment.

Our key insight is that smell information is only **partially observable** from visual input. While some odor sources correspond to visible objects, many olfactory factors arise from contextual or latent environmental conditions (e.g., traffic emissions, ventilation, nearby activities). An example is shown in Figure 1: a second view of a scene reveals olfactory context outside the first view; often, at training, we only have access to the first view. Consequently, direct

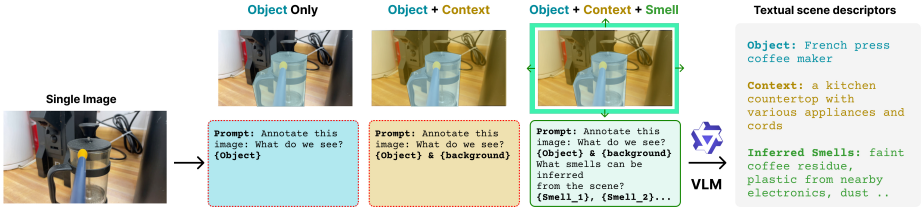


Fig. 2: Textual scene descriptors via VLM inference. Given an image, a pretrained VLM generates object, contextual, and inferred smell descriptors, which serve as semantic guidance for olfactory representation learning.

visual–smell correspondence provides incomplete supervision for learning olfactory representations.

To address this limitation, we introduce language to fill the information gap between visual observation and smell measurements. Vision–Language models (VLMs) [29, 46] can infer plausible environmental conditions and latent factors from visual context, producing textual descriptors T that expose potential olfactory cues. In this paper, our main idea consists of leveraging these inferred cues to learn olfactory representations.

Our proposed framework therefore combines three components as illustrated in Figures 2-3: (1) VLM-based semantic scene augmentation that produces descriptors T (Section 3.1); (2) multimodal olfactory representation learning that aligns z^S with visual and textual embeddings (Section 3.2); and (3) latent smell disentanglement that separates object-specific and contextual odor components (Section 3.3).

3.1 VLM-based Semantic Scene Augmentation

The available data provide images I and smell measurements $\mathbf{X}$, but do not explicitly describe the environmental factors that contribute to the observed olfactory signal. To expose such latent smell information, we use a pretrained Vision-Language model (VLM) to generate textual scene descriptors.

Given an image I, the VLM produces a set of descriptors:

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

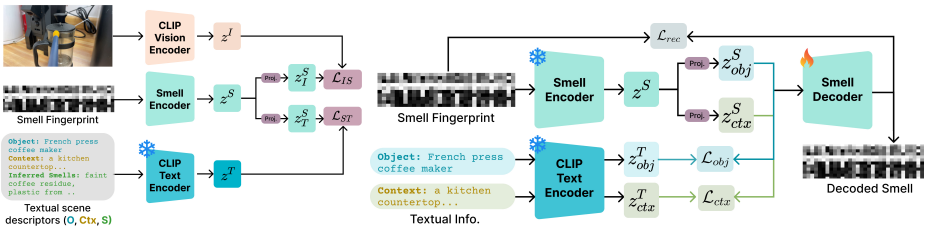
where t_{obj} denotes a textual description of the primary object in the scene, t_{ctx} describes the surrounding environmental context, and $\{t^{(k)}_{\text{smell}}\}_{k=1}^{K}$ are textual descriptors referring to plausible ambient smell cues suggested by the scene context. These descriptors provide 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 encourages the VLM to infer potential olfactory elements.

These descriptors are combined and prompted into a long textual description and encoded using the frozen CLIP [39] text encoder: $z^T = f_T(\text{prompt}(T)) \in \mathbb{R}^{512}$.

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.

Annotation quality. To verify that VLM-generated descriptors infer real olfactory context rather than plausible hallucinations, we use a VLM judge given a held-out second view of the same scene (View 2, not seen during annotation; see Figure 1 for an example). The judge scores each annotation on two criteria: **plausibility** (contextually reasonable given both views?), reaching **98.6%**; and **View-2 confirmation** (specifically predicts 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. A visual illustration of the evaluation setup is provided in the supplementary material.

3.2 Learning Multi-modal Olfactory Representations



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

Fig. 3: Overall framework for multimodal olfactory representation learning and disentanglement. (Step 2) 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_I^S and z_T^S. (Step 3) Latent Smell Disentanglement: Leveraging the pretrained encoder, we disentangle the smell representation into object (z_{obj}^S) and context (z_{ctx}^S) components. These components are aligned with their respective linguistic descriptors via $\mathcal{L}_{\text{obj}}$ and $\mathcal{L}_{\text{ctx}}$.

With visual observations and augmented smell descriptors available, we now learn an olfactory representation $z^S \in \mathbb{R}^d$ from Equation 1 that is aligned with both visual and language modalities.

Olfactory Encoder. The smell signal $\mathbf{X}$ is processed using a Transformer-based encoder that models temporal dependencies across sensor measurements. The resulting embedding z^S captures the global, both visual and language, structure of the olfactory signal.

To enable alignment with visual and textual modalities, we apply two modality-specific projection heads

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

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

Visual and Textual Encoders. We obtain visual embeddings $z^I = f_I(I) \in \mathbb{R}^{512}$ using the CLIP image encoder, and textual embeddings z^T from the CLIP representation of the VLM-generated descriptors as described in Section 3.1. While the text encoder remains frozen to preserve semantic structure, we allow the image encoder to be fine-tuned during the alignment stage. This enables the visual backbone to adapt its feature extraction to prioritize subtle visual cues, such as material textures or environmental conditions, that are correlated with olfactory signatures but may be secondary or typically overlooked in general-purpose visual representations (for instance, kitchen spices are often disregarded when it comes to captioning, even though they reveal smell information).

Multimodal Alignment. We train the smell encoder by aligning the projected smell representations with their corresponding visual and textual embeddings using symmetric InfoNCE losses [36, 39]. 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

$$\mathcal{L}_{\text{total}} = \lambda_{IS}\mathcal{L}_{IS} + \lambda_{ST}\mathcal{L}_{ST} \quad . \tag{2}$$

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 originating from the target object as well as background environmental factors. For example, the smell measurement of a particular object undoubtedly captures also smell information about the background. To separate the contribution of object vs context background, we introduce a latent decomposition of the olfactory representation. 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 'scene' it is located. During training, each decoded smell component is aligned with its corresponding textual descriptor 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} (Section 3.1).

Denoting z^T_{obj} and z^T_{ctx} the corresponding CLIP text embeddings. We enforce this decomposition through contrastive losses. Let $\text{sim}(\cdot,\cdot)$ 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 three losses:

$$\mathcal{L}_{\text{dis}} = \mathcal{L}_{\text{obj}} + \mathcal{L}_{\text{ctx}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} \quad . \tag{3}$$

λ_{rec} is a weighting factor for the reconstruction loss, which acts as a regularizer to prevent embedding collapse. This decomposition encourages the model 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 [38], the largest multimodal benchmark for in-the-wild olfactory perception. NYS contains 7,000 paired image-smell samples across 3,500 object categories. Each olfactory sample comprises a 32-channel resistance signal from an e-nose array, consisting of a baseline recording ($\mathbf{B} \in \mathbb{R}^{T\times 32}$) of ambient air and a sample recording ($\mathbf{S} \in \mathbb{R}^{T\times 32}$) taken near the object, where T denotes the temporal dimension (typically 14 timesteps). Following [38], we use a unified olfactory input $\mathbf{X}$ formed by the temporal concatenation of the two signals: $\mathbf{X} = [\mathbf{B};\mathbf{S}] \in \mathbb{R}^{2T\times 32}$. We use 5,996 samples for training and 936 for validation. All baselines and ablations are evaluated on the same split.

Implementation Details. Our olfactory encoder f_S is implemented as a Transformer with $L = 6$ layers and $H = 8$ attention heads. The input signal $\mathbf{X} \in \mathbb{R}^{2T \times 32}$ is projected to a $d = 448$ dimensional embedding and augmented with sinusoidal positional encodings. We map the resulting olfactory representation z^S to the 512-dimensional CLIP visual and textual spaces using two MLP projection heads. For semantic augmentation, we use Qwen3VL-30B [46] to generate descriptors $\{t_{\text{obj}}, t_{\text{ctx}}, t_{\text{smell}}\}$. We employ CLIP ViT-B/16 [39] 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.

Fusion for Joint-modality Retrieval. To perform joint retrieval, we evaluate two distinct fusion paradigms to unify the image-aligned (z_I^S) and textual-aligned (z_T^S) smell embeddings.

1. **Latent-level Fusion**: Our primary method, which constructs a single blended representation $\tilde{z}$ before computing similarity:

$$\tilde{z}_i = \frac{\alpha\, z_{I,i}^S + (1-\alpha)\, z_{T,i}^S}{\left\| \alpha\, z_{I,i}^S + (1-\alpha)\, z_{T,i}^S \right\|_2}$$

   The retrieval score is then calculated by comparing this unified vector to both targets:

$$\text{score}_{ij} = \alpha \cdot \text{sim}(\tilde{z}_i, z_j^I) + (1-\alpha) \cdot \text{sim}(\tilde{z}_i, z_j^T)$$

2. **Similarity-level Fusion**: A late-fusion approach where the joint score is a weighted combination of independent similarity scores:

$$\text{score}_{ij} = \alpha \cdot \text{sim}(z_{I,i}^S, z_j^I) + (1-\alpha) \cdot \text{sim}(z_{T,i}^S, z_j^T)$$

Table 1: Impact of semantic complexity 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 object (**O**), context (**Ctx**), and inferred ambient smells (**I**) yields the most robust cross-modal alignment across both single and joint retrieval tasks.

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)

Latent-level fusion forces the model to find a single point in the latent manifold that satisfies both visual and linguistic constraints simultaneously. Unless otherwise specified, we use Latent-level fusion for all joint retrieval experiments, with the fusion weight α selected via grid search on the validation set.

Baselines. As the current state-of-the-art on the NYS dataset [38] 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 multi-modal 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 [39] 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

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 provides a critical semantic bridge that visual features alone cannot offer.

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 successfully outperforms the reproduced NYS baseline (R@5 from 20.0 to 22.1) even though it uses only 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 learned 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.

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

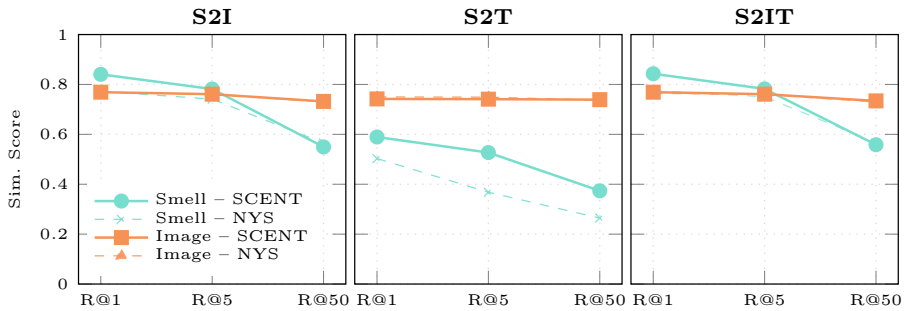


Fig. 4: Mean cosine similarity between the ground-truth and top-k retrieved items, in frozen CLIP image space (orange) and learned olfactory space (teal), for $k \in \{1, 5, 50\}$. Solid lines: SCENT; dashed: NYS.

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

In this section, we present our ablation studies. We focus on two primary axes: the granularity of the language supervision (Table 2) and the strategy for multimodal feature fusion (Table 4).

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 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 semantic supervision, we apply the Image Bridge protocol to SCENT: we disable the ST head and route S2T/S2IT retrieval via the IS

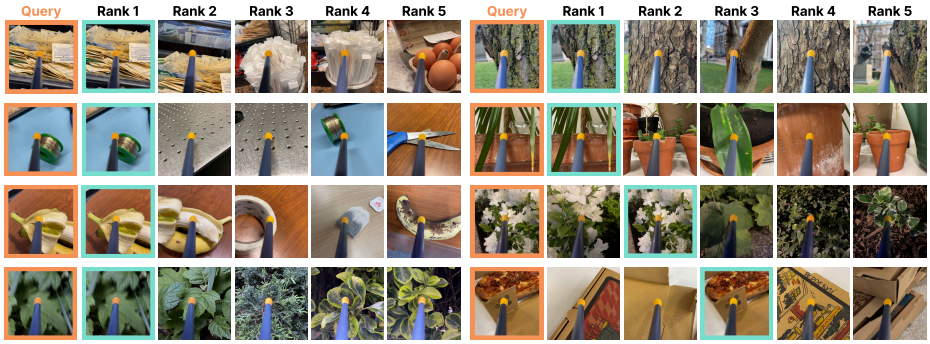


Fig. 5: Qualitative results for S2IT retrieval. For a given olfactory query, we show the ground-truth image in the **Query** column and the top-5 images from the retrieved (Image, Text) pairs. The **Correct** retrieval is highlighted. SCENT demonstrates robust cross-modal alignment across diverse categories, 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).

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.

Fusion Strategies. We evaluate the sensitivity of our joint retrieval (S2IT) to different fusion paradigms and modality weightings (a). As shown in Table 4, both Similarity-level and Latent-level fusion exhibit a similar upward trend as the visual weighting a increases. While Similarity-level fusion provides slightly better performance at lower visual weights, Latent-level fusion demonstrates a more significant performance gain as the modalities are balanced, reaching an R@5 of 22.9. This suggests that while Similarity fusion acts as a "greedy" ensemble of independent heads, our chosen Latent-level fusion succeeds in forcing the olfactory embedding into a singular, semantically coherent point in the CLIP manifold that satisfies both visual and textual constraints simultaneously.

4.5 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

Table 2: State-of-the-art comparison on the NYS dataset. We compare SCENT against the vision-only NYS baseline [38] 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* [38]				—	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 [38].
† Results reproduced.

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. SCENT with IB still outperforms NYS† (adapted) with IB, showing that co-training with the ST loss also improves the IS head.

Method	IB	S2T		S2IT	
		R@5	R@20	R@5	R@20
NYS† (adapted)	✓	8.1	19.2	20.0	40.9
SCENT (ours)	✓	9.1	18.8	21.4	43.8
SCENT (ours)	✗	**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.

Table 4: Ablation of fusion strategies and modality weighting. We evaluate the sensitivity of joint-modality retrieval (S2IT) to different fusion paradigms and visual-textual weighting coefficients (a). Results are reported using the full semantic template (**O** + **Ctx** + **S**).

Fusion	a	R@1	R@5	R@10	R@20
Similarity-level	0.3	5.0	20.3	32.5	44.3
	0.5	5.8	**22.1**	32.2	44.1
	0.7	**6.3**	22.0	**34.0**	**44.6**
Latent-level	0.3	4.7	17.5	27.6	40.7
	0.5	6.1	20.8	31.2	42.9
	0.7	**6.2**	**22.9**	**32.9**	**42.9**

Table 5: Recombination retrieval for olfactory disentanglement. We evaluate the model's ability to synthesize novel olfactory fingerprints by recombining object and context factors that were seen during training only in isolation. **Decoded Synthesis** (ours) consistently outperforms the **Raw Recombination** baseline, demonstrating that our latent decomposition successfully captures the semantic essence of scent components rather than relying on raw signal patterns.

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

Quantitative Analysis. As shown in Table 5, we evaluate the retrieval performance of these synthetic queries within our latent space, by passing the query smell fingerprint **X** through our smell encoder. Our Decoded Synthesis outperforms the Raw Recombination baseline. While the raw baseline plateaus at 9.8 Recall@20, our synthesized fingerprints achieve 13.7 Recall@20.

5 Limitations and Conclusion

Despite the performance gains of SCENT, several challenges remain. The reliance on metal-oxide (MOX) sensors introduces inherent stochasticity, as these hardwares are prone to temporal drift and environmental cross-sensitivity to humidity and temperature. Furthermore, while the NYS dataset is a significant milestone, a substantial "data gap" remains between olfactory benchmarks and the web-scale datasets used to train visual-linguistic 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 as a semantic bridge, leveraging the extensive world knowledge embedded in VLMs to infer plausible ambient olfactory cues. By implementing 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 significantly enhances cross-modal retrieval performance across both single and joint modality tasks. Furthermore, our language-guidance mechanism achieves superior decompositionality of olfactory information, facilitating improved generalization to unseen scenarios and robust performance on out-of-distribution (OOD) samples.

References

1. Abeßer, J.: A review of deep learning based methods for acoustic scene classification. Applied Sciences **10**(6), 2020 (2020)
2. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: NeurIPS (2016)
3. Bender, T., Sørensen, S., Kashani, A., Eldjarn Hjorleifsson, K., Hyldig, G., Hauberg, S., Belongie, S., Warburg, F.: Learning to taste: A multimodal wine dataset. Advances in Neural Information Processing Systems **36**, 7351–7360 (2023)
4. 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)
5. Chen, H., Huo, D., Zhang, J.: Gas recognition in e-nose system: A review. IEEE transactions on biomedical circuits and systems **16**(2), 169–184 (2022)
6. 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)
7. Cicchetti, G., Grassucci, E., Comminiello, D.: A triangle enables multimodal alignment beyond cosine similarity. In: NeurIPS (2025)
8. Cicchetti, G., Grassucci, E., Sigillo, L., Comminiello, D.: Gramian multimodal representation learning and alignment. In: ICLR (2025)
9. De Sa, V.: Learning classification with unlabeled data. In: NeurIPS (1993)
10. Debnath, T., Nakamoto, T.: Predicting human odor perception represented by continuous values from mass spectra of essential oils resembling chemical mixtures. PloS one **15**(6), e0234688 (2020)
11. Dou, Y., Yang, F., Liu, Y., Loquercio, A., Owens, A.: Tactile-augmented radiance fields. In: CVPR (2024)
12. Elizalde, B., Deshmukh, S., Al Ismail, M., Wang, H.: CLAP: learning audio concepts from natural language supervision. In: Proc. ICASSP (2023)
13. Erlangga, F., Wijaya, D.R., Wikusna, W.: Electronic nose dataset for classifying rice quality using neural network. In: 2021 9th International Conference on Information and Communication Technology (ICoICT). pp. 462–466. IEEE (2021)
14. 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. Advances in Neural Information Processing Systems **37**, 89098–89124 (2024)
15. Feng, D., Dai, W., Li, C., Pernigo, A., Wen, Y., Liang, P.P.: Smellnet: A large-scale dataset for real-world smell recognition. arXiv preprint arXiv:2506.00239 (2025)
16. Feng, D., Dai, W., Li, C., Pernigo, A., Wen, Y., Liang, P.P.: Smellnet: A large-scale dataset for real-world smell recognition. arXiv preprint arXiv:2506.00239 (2025)
17. Feng, S., Zhu, X.: A transformer-based deep learning network for underwater acoustic target recognition. IEEE Geoscience and Remote Sensing Letters **19**, 1–5 (2022)

18. Fundurulic, A., Faria, J.M., Inácio, M.L.: Advances in electronic nose sensors for plant disease and pest detection. Engineering Proceedings **48**(1), 14 (2023)
19. 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 **3**(3), 449–469 (2023)
20. Gao, Y., Hendricks, L.A., Kuchenbecker, K.J., Darrell, T.: Deep learning for tactile understanding from visual and haptic data. In: 2016 IEEE international conference on robotics and automation (ICRA). pp. 536–543. IEEE (2016)
21. 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 **17**(2), 026006 (2023)
22. 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)
23. 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)
24. 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)
25. 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 **355**(6327), 820–826 (2017)
26. 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
27. 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 **381**(6661), 999–1006 (2023)
28. Lee, M.A., Zhu, Y., Srinivasan, K., Shah, P., Savarese, S., Fei-Fei, L., Garg, A., Bohg, J.: Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In: 2019 International conference on robotics and automation (ICRA). pp. 8943–8950. IEEE (2019)
29. 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
30. 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)
31. 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 **47**(2), 708–724 (2024)
32. 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)
33. 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 **508**, 293–304 (2022)
34. 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 **115**, 593–606 (2019)
35. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y., et al.: Multimodal deep learning. In: ICML (2011)
36. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
37. Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E.H., Freeman, W.T.: Visually indicated sounds. In: CVPR (2016)
38. 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)
39. 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)
40. 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 **588**(7836), 118–123 (2020)
41. Ruan, L., Hu, A., Song, Y., Zhang, L., Zheng, S., Jin, Q.: Accommodating audio modality in clip for multimodal processing. In: AAAI (2023)
42. 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)
43. Saporta, A., Puli, A.M., Goldstein, M., Ranganath, R.: Contrasting with symile: Simple model-agnostic representation learning for unlimited modalities. In: NeurIPS (2024)
44. Snitz, K., Yablonka, A., Weiss, T., Frumin, I., Khan, R.M., Sobel, N.: Predicting odor perceptual similarity from odor structure. PLoS computational biology **9**(9), e1003184 (2013)
45. 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 **4**, 104–115 (2020)
46. Team, Q.: Qwen3 technical report (2025), `https://arxiv.org/abs/2505.09388`
47. 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 **20**(23), 14302–14309 (2020)
48. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3156–3164 (2015)
49. 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 **45**(10) (2025)
50. 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 (2024)
51. Wilson, A.D.: Review of electronic-nose technologies and algorithms to detect hazardous chemicals in the environment. Procedia Technology **1**, 453–463 (2012)
52. Xiao, J., Yao, A., Li, Y., Chua, T.S.: Can i trust your answer? visually grounded video question answering. In: CVPR (2024)
53. 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)

54. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27091–27101 (2024)
55. Yang, F., Ma, C., Zhang, J., Zhu, J., Yuan, W., Owens, A.: Touch and go: Learning from human-collected vision and touch. arXiv preprint arXiv:2211.12498 (2022)
56. Yuan, W., Wang, S., Dong, S., Adelson, E.: Connecting look and feel: Associating the visual and tactile properties of physical materials. In: CVPR (2017)
57. 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)
58. 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)
59. Zhao, Y., Misra, I., Krähenbühl, P., Girdhar, R.: Learning video representations from large language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6586–6597 (2023)
60. 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)