symmetry

MDPI

Article

Synthesizing Olfactory Understanding: Multimodal Language Models for Image–Text Smell Matching

Sergio Esteban-Romero *, Iván Martín-Fernández, Manuel Gil-Martín and Fernando Fernández-Martínez

Grupo de Tecnología del Habla y Aprendizaje Automático (THAU Group), Information Processing and Telecommunications Center, E.T.S.I. de Telecomunicación, Universidad Politécnica de Madrid (UPM), 28040 Madrid, Spain; ivan.martinf@upm.es (I.M.-F.); manuel.gilmartin@upm.es (M.G.-M.); fernando.fernandezm@upm.es (F.F.-M.)

* Correspondence: sergio.estebanro@upm.es; Tel.: +34-910672033

Abstract

Olfactory information, crucial for human perception, is often underrepresented compared to visual and textual data. This work explores methods for understanding smell descriptions within a multimodal context, where scent information is conveyed indirectly through text and images. We address the challenges of the Multimodal Understanding of Smells in Texts and Images (MUSTI) task by proposing novel approaches that leverage language-specific models and state-of-the-art multimodal large language models (MM-LLMs). Our core contribution is a multimodal framework using language-specific encoders for text and image data. This allows for a joint embedding space that explores the semantic symmetry between smells, texts, and images to identify olfactory-related connections shared across the modalities. While ensemble learning with language-specific models achieved good performance, MM-LLMs demonstrated exceptional potential. Fine-tuning a quantized version of the Qwen-VL-Chat model achieved a state-of-the-art macro F1-score of 0.7618 on the MUSTI task. This highlights the effectiveness of MM-LLMs in capturing task requirements and adapting to specific formats.

Keywords: olfactory understanding; multimodal perception; Contrastive Language–Image Pretraining (CLIP); Multimodal Large Language Models (MM-LLMs)

Received: 6 June 2025
Revised: 23 July 2025
Accepted: 15 August 2025
Published: 18 August 2025

Citation: Esteban-Romero, S.; Martín-Fernández, I.; Gil-Martín, M.; Fernández-Martínez, F. Synthesizing Olfactory Understanding: Multimodal Language Models for Image–Text Smell Matching. *Symmetry* **2025**, *17*, 1349. https://doi.org/10.3390/sym17081349

1. Introduction

Despite its profound influence on memory and emotion [1], the sense of smell remains understudied compared to the rapid advancements in Natural Language Processing (NLP) and Computer Vision (CV) [2,3]. Unlike other sources of information such as vision, where wavelength maps to color, or audition, where frequency maps to pitch, olfaction is based on complex interactions between molecules and olfactory receptors, which poses a poorly understood correlation between chemical structures and olfactory percepts [4–6].

Humans can recognize or imagine smells from visual cues or hearing a description of an object, reflecting a natural form of cross-modal learning. Inspired by this ability, recent models integrate visual and olfactory data to enable cross-modal recognition and understanding [7]. Building on this, smells exhibit an inherent capacity to evoke memories or emotions, resulting in efforts such as the Multimodal Understanding of Smells in Texts and Images (MUSTI) challenge held at MediaEval2023 [8,9]. Specifically, we delve into Subtask 1 of MUSTI, which focuses on predicting whether image–text pairs from the 17th to the 20th centuries (in various languages, including English, German, Italian, and French) contain references to the same smell sources or not.

Many existing models that integrate visual and textual components concentrate on tasks such as entailment or captioning, which not only provide a strong grasp of semantic similarity but also emphasize the importance of semantic symmetry between both modalities. Although these models excel at understanding general concepts and associations between images and texts, their broad focus necessitates fine-tuning to adapt their understanding to highly specific tasks such as olfactory recognition. In this domain, the unique characteristics and associations are not readily captured by generic multimodal models. Additionally, most pretrained models are trained using photograpic data, which further supports the importance of adaptation in a scenario where images correspond to paintings [3,10]. Thus, fine-tuning becomes indispensable for these models to properly discern and interpret olfactory sources accurately, leveraging their inherent capabilities in multimodal understanding.

In our previous work [11], we employed language-specific models based on the Contrastive Language–Image Pretraining (CLIP) framework [12] to address the multilingual nature of the MUSTI dataset. However, traditional CLIP implementations struggled to capture the specific relationships between smells, texts, and images. To overcome this limitation, we adopted a supervised training methodology to represent both modality embeddings in a multimodal space that effectively links the textual and visual representations but conditioned to the proper identification of smells. We further enhanced our approach by introducing an ensemble method that takes advantage of the combined strengths of multiple models. This technique utilizes a simple linear layer to optimize the weights assigned to each model's predictions, resulting in a more robust final outcome. Building upon our previous work, this paper delves deeper into the potential of multimodal large language models (MM-LLM) as olfactory experts. Specifically, we investigate the use of Qwen-VL-Chat models [13] in their pretrained and fine-tuned states to enhance olfactory understanding. The fine-tuning process involves the low-rank adapter (LoRA) technique to achieve an efficient adapted model while fewer computational resources are required.

Finally, our work also contributes to the field by systematically evaluating the impact of different class imbalance mitigation techniques on the performance of MM-LLMs for olfactory understanding. By comparing the performance of models trained on undersampled negative, balanced, and unbalanced positive class datasets, we provide valuable insights into effective strategies for addressing class imbalance in this specific task. This knowledge can be valuable for researchers working with similar datasets with skewed class distributions.

This research presents the following contributions:

- We propose a method to model semantic symmetry between images and texts by aligning their representations through shared olfactory references. This enables improved detection of smell-related elements in multimodal data using a multilingual ensemble of language-specific models.
- We explore the potential of fine-tuning multimodal large language models (MM-LLMs), particularly Qwen-VL-Chat, for significantly improved olfactory recognition in the MUSTI task.
- We conduct a systematic evaluation of class imbalance mitigation strategies, including majority-class downsampling, to address the pronounced label skew in the MUSTI dataset and improve model robustness.

The structure of the paper is as follows. Section 2 offers an overview of the related work in the field of olfactory matching, focusing on the approaches that have supported the solutions developed to identify whether text–image pairs contain references to similar scent sources. Furthermore, the latest approaches presented at MediaEval 2023 [8] in the MUSTI task are commented on. Section 3 provides an overview of the dataset used in this study. Section 4 describes the solutions carried out, covering CLIP-based language-specific

models, an ensemble approach that combines the individual expertise of these models, and the use of MM-LLMs to understand the semantics present in smells that enhance olfactory matching. The experimental setup, along with the results obtained for each of the approaches described above, are presented in Section 5. A summary of the key findings obtained throughout our experimentation alongside future studies to be performed based on them is provided in Section 6.

2. Related Work

2.1. Multimodal Image–Text Alignment Overview

Multimodal learning has advanced significantly in recent years, and numerous architectures have been developed to model the existing relationships between different modality sources containing references to similar elements. Building on this foundation, works like VirTex [14] explore the benefits of natural language supervision for multimodal training using Convolutional Neural Networks (CNNs) as image encoders and language models, demonstrating improved performance in downstream tasks due to the resulting semantic richness while reducing the scale of training sets required. Similarly, ConVIRT [15] enhances visual–textual representation alignment by using contrastive learning to distinguish true image–text pairs from randomly paired examples. It jointly trains a text encoder and an image encoder to learn a shared embedding space in which representations of matching image–text pairs are drawn closer together. A central approach in this area is Contrastive Language–Image Pretraining (CLIP) [12], a simplified extension of ConVIRT, trained on a newly collected dataset of 400 million image–text pairs. The scale of this dataset enables CLIP to achieve strong performance on a wide range of tasks, even in zero-shot settings.

2.2. Multimodal Image–Text Alignment Based on Smell Sources

In this work, we address the specific challenge of aligning image and text modalities, conditioned on the presence of references to smell sources in both. Within the framework of the MUSTI challenge, presented at MediaEval 2023 [16], several novel approaches were introduced to model olfactory relationships between paintings and their associated multilingual textual descriptions. Akdemir et al. [17] evaluated the performance of fine-tuning two state-of-the-art models, VilBERT [18] and mUNITER [19], to handle multiple languages present in the dataset, reporting a macro F1-score result of 0.6176. The proposed setup addressed the task as a visual entailment problem, where the image serves as the premise and the text as the hypothesis. The model must classify the relationship between them as entailment, neutral, or contradiction [20]. In this context, an entailment prediction indicates a positive relationship, i.e., both modalities reference the same smell sources, while neutral or contradiction predictions are treated as negative. Other entries in the MUSTI challenge explored approaches such as directly utilizing CLIP embeddings [21] to later determine whether a pair is related based on the cosine similarity of its representations achieving a 0.6176 macro F1-score. Ngoc-Duc et al. [22] obtained image representations using a Vision Transformer (ViT) [23] and Resnet-34 [24] models and BERT [25] as the text encoder. They incorporated an Image-Guided Feature Extractor (IGFE) to enhance alignment between the image and text embeddings by injecting image features into the textual representation. This allows the text backbone to attend more effectively to visual cues relevant to olfactory references. Using this approach, they achieved a macro F1-score of 0.7442 on the test set. Previous MUSTI challenge entries highlight the limitations of directly using pretrained models for olfactory matching as these models have not been specifically trained to capture the nuances of smell descriptions and relate them to visual features despite their demonstrated capabilities when applied to general-domain scenarios [12,17].

2.3. Multimodal Large Language Model Overview

In recent years, state-of-the-art large language models [26] (LLMs) have revolutionized the field of Natural Language Processing (NLP), demonstrating remarkable capabilities in understanding texts and solving numerous tasks. Many such models exist, such as LLaMa [27], Gemma [28], and Qwen [29], among many others. These models are based on the transformer [30] architecture and have demonstrated outstanding performance in several tasks and domains due to their large-scale pretraining.

To further expand their capabilities, recent advances have enabled LLMs to integrate and process multiple data modalities, including images, audio, etc., broadening the range of domains they can effectively address [31–33]. These multimodal large language models (MM-LLMs) leverage architectures that enable LLMs to process different sources of information via projecting out-of-domain modalities into their latent space, benefiting from their semantic understanding and reasoning capabilities to promote greater understanding of the existing relationships between them. For example, when working with text and image pairs, the Qwen-VL [13] model enables such alignment by incorporating a vision encoder that extracts features from the images that are later projected into the semantic space of the LLM, facilitating seamless integration and interaction between visual and textual data.

2.4. Multimodal Large Language Models for Smell Identification

The potential of MM-LLMs was also explored in the MUSTI challenge. Srinivasan et al. [34] leveraged a BLIP-based [35] MM-LLM, in a frozen and fine-tuned setup, to generate textual descriptions of images for comparison with the corresponding texts associated with each image. In particular, they reported macro F1-scores of 0.5591 and 0.4893 for the fine-tuned and base models, respectively.

Furthermore, the current state of the art for the binary classification task proposed in MUSTI was demonstrated in the work of Kurfalı et al. [36], who fine-tuned the Large Language and Vision Assistant (LLaVa) [37] model using a resource-efficient technique called low-rank adaptation (LoRA) [38] on the 13B parameter version of the model. Their approach obtained a macro F1-score of 0.7760. In their setup, prompts are used to guide the multimodal large language model (MM-LLM) in determining the olfactory relationship between image–text pairs, with the model required to respond directly with 'YES' or 'NO'. They also explored a zero-shot setting to assess the pretrained model's capabilities using an in-house development set that represents only 10% of the full dataset. Both the 7B and 13B parameter versions of LLaVa were evaluated under these conditions, reporting significant improvements when fine-tuning but not when increasing the parameter count.

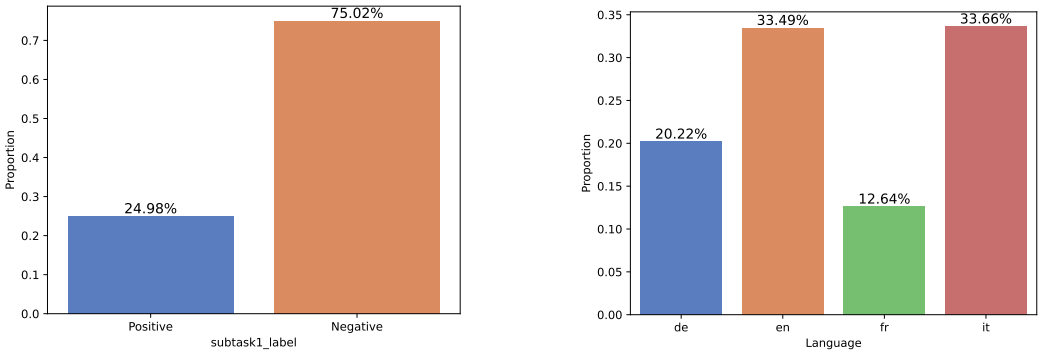
These results highlight the effectiveness of MM-LLMs even without extensive fine-tuning for the MUSTI task. Another example is the Qwen-VL-Chat model [13], which is also available in a quantized version. This model was primarily trained on caption data and dialogue data generated through LLM self-instruction, mainly in English and Chinese. It has demonstrated strong capabilities in fine-grained visual understanding and grounding, along with its proven success on various vision and language benchmarks.

3. Data

The dataset used in this work was released for the MUSTI challenge at MediaEval 2022. A comprehensive description of the dataset is provided in [8]. The original split consists of 2374 image–text pairs for training and 814 for testing. It consists of multilingual copyright-free texts, gathered from historical books and documents, and images from the 17th to the 20th centuries. The images have been collected from RKD, Bildindex der Kunst und Architektur, Museum Boijmans, Ashmolean Museum Oxford, and Plateforme ouverte du patrimoine, annotated with over 80 categories related to smell, including objects and gestures

like flowers, food, animals, and sniffing and holding the nose [8]. The detailed description of the taxonomy used for annotating such datasets can be found in the description of the Odoeuropa project [1].

A significant challenge of the dataset is the class imbalance in the training data: nearly 75% of the pairs show a negative relationship in terms of smell references. This imbalance can hinder the ability of models to learn the specific features that indicate matching olfactory descriptions in image–text pairs. Although less remarkable, there is also an imbalance with respect to the different languages available, which are English (en), French (fr), Italian (it), and German (de). Figure 1 illustrates the distribution of the examples in terms of both dimensions discussed.



(**a**) Distribution of examples in terms of the target label

(**b**) Distribution of examples regarding different languages in texts

Figure 1. Distributions of MUSTI dataset in terms of target label (**a**) and language in text (**b**).

To address the class imbalance in the MUSTI dataset (75% negative class, as shown in Figure 1a), we created three experimental setups to examine their impact on model performance and generalizability, see Table 1. Specifically, we aimed to reduce potential biases caused by the imbalanced negative class while maintaining the distribution of languages within the dataset. We achieved this by under-sampling the negative class, ensuring the other classes remained proportionally represented. Additionally, we evaluated a balanced scenario where the training data contained an equal number of positive and negative examples. Finally, we explored an unbalanced positive class setup to see if prioritizing positive examples during training, even with a reduced dataset size, could improve the classification of positive relationships between image–text pairs.

Table 1. Distribution of examples for each of the different experimental setups considered with the mean value of the cosine similarity for each image–text pair on each test set.

Exp. Setup	Total	Pos.	Neg.	Test
Unbal. Neg.	2374	593	1781	356 (15%)
Balanced	1218	593	625	122 (10%)
Unbal. Pos.	994	593	401	146 (15%)

4. Approach

4.1. Language-Specific Models

First, we propose a solution based on the simultaneous fine-tuning of a textual and a visual branch composed of a text encoder and an image encoder followed by a projection layer with the aim of learning a joint embedding space, where representations containing references to the same smell sources are fostered for both text and image data in pairs labeled as *YES* (matching smell), while representations for *NO* pairs are kept distinct.

However, although CLIP's approach of learning similar representations for matched text–image pairs is appealing [12], it is not directly applicable to our imbalanced dataset due to its underlying assumptions about the data. The standard CLIP implementation is designed and particularly well suited for datasets where all input pairs are inherently aligned and positive (the text describes the image). This assumption translates to a loss function that maximizes similarity for these positive pairs (diagonal elements in a large similarity matrix) while minimizing similarity for all other pairings. Therefore, this strategy becomes inefficient for our task for the following two main reasons:

- Imbalanced Data: Our dataset is heavily skewed toward negative examples (i.e., no matching smell). Applying the CLIP training scheme in this context would require discarding a substantial portion of the training data, approximately 75%, as it is designed to align representations between positive pairs while pushing non-matching pairs apart. As a result, negative relationships would be excluded from training, limiting the model's ability to learn from the full dataset.
- Captioning Design Mismatch: CLIP assumes a strong correlation between images and their corresponding captions, where each caption directly describes the image or references its elements or context. However, in our case, this assumption does not hold. Even for positive pairs, the text may not explicitly describe the smell depicted in the image as it can consist of passages that refer to olfactory concepts without directly relating to the visual content.

Therefore, CLIP's core assumptions and loss function are suboptimal for our task of identifying matching smell sources in text–image pairs with a significant negative class. We need a model that can effectively utilize all available data, including negative examples, and capture the more nuanced relationships between text and image in the context of smell understanding.

To overcome these limitations, we propose end-to-end model training using a supervised approach, as illustrated in Figure 2. In this setup, models are trained to directly predict one of the provided labels, matching or non-matching smell, rather than relying solely on self-supervised learning. Nevertheless, we retain a contrastive learning objective aimed at maximizing agreement between positive examples (matching smell) and minimizing agreement between negative ones (non-matching smell) in the latent space. This is feasible because, as noted previously, a single image can be paired with multiple text descriptions, some of which may be positive and others negative. As a result, we expect the model to structure the latent embedding space in a manner similar to that of a purely self-supervised contrastive learning approach.

The loss function utilized is the binary cross-entropy (BCE) loss, which compares the model's predictions (obtained via cosine similarity) with the actual labels of the training data. Cosine similarity quantifies the similarity between vectors, with a value of 1 indicating perfect similarity and -1 representing complete dissimilarity. In our case, we normalize the cosine similarity values between 0 and 1 before applying a fixed classification threshold. As commented, it is expected that the model learns to project both modality representations into a embedding space, where a cosine similarity close to 1 represents image–text pairs with matching smell references, while a value close to 0 indicates no similarity.

For the vision branch, we used a Vision Transformer (ViT) (https://huggingface.co/google/vit-large-patch16-224-in21k, accessed on 16 August 2025) [23] pretrained on ImageNet-21k [39]. To take advantage of the information from the different languages in the dataset, we explored a variety of text encoders pretrained on language-specific data according to those present in the dataset: English MPNET (https://huggingface.co/sentence-transformers/all-mpnet-base-v2, accessed on 16 August 2025) [40], French CamemBERT (https://huggingface.co/dangvantuan/sentence-camembert-large, accessed on 16 August

2025) [41], Italian BERT (https://huggingface.co/dbmdz/bert-base-italian-uncased, accessed on 16 August 2025), and German BERT (https://huggingface.co/bert-base-german-dbmdz-uncased, accessed on 16 August 2025). All model checkpoints are available at *huggingface*.

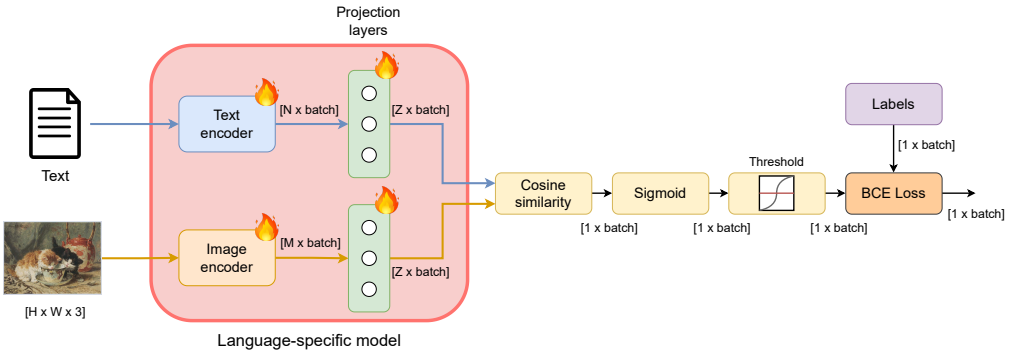


Figure 2. Architecture of the language-specific models proposed with corresponding output dimensions. M and N represent the output dimensions of the image and text encoder, respectively. Z represents the output dimension of the projection layer after each encoder, which must be the same to obtain the joint embedding space. After cosine similarity is computed, a vector of length equal to batch size is passed through the rest of the modules. The flame represents the parts fine-tuned when training and the snowflake those that are kept frozen.

For both the visual and textual branches, we use the [CLS] token output as the modality-specific representation, which is a learnable embedding that serves as a global summary of the input in BERT- [25] and ViT-based [23] models.

Due to the limited size of the dataset and the scarcity of examples for some languages, we opted to train the text encoders on data from all four languages (English, French, Italian, and German) rather than using separate pretrained models for each language. Although this approach may be considered suboptimal, we hypothesize that linguistic similarities among some of the languages could potentially benefit underrepresented languages such as French, which only represents 12.54% of the data. Also note that this approach, while leveraging the full dataset, also results in the creation of language-dependent visual encoders because the shared ViT model in the vision branch adapts its representation based on the specific language in which it has been fine-tuned during text–image joint training.

To obtain a suitable joint embedding space where image and text representations can be directly compared based on the olfactory references they contain, both encoder outputs must be projected into a common dimensionality, which leads to the incorporation of the projection layer after each encoder. The objective of the projection layer is to represent the encoder output with embeddings of a common size (which is set to 256 in our case), denoted as Z in Figure 2. This step is crucial because the comparison between both representations via cosine similarity is not possible otherwise. Additionally, it enables the use of text encoders with different latent representation dimensionality as each projection layer will be tailored to either compress or expand them to the specified hidden dimension size. The projection layer is a Multi-Layer Perceptron (MLP) formed by two linear layers with layer normalization and a dropout of 0.3 in the last one.

The distinction on whether the relationship between the encoded representations for both modalities is positive or negative is computed in the resulting multimodal "smell embedding space" using cosine similarity (a measure of similarity between vectors) between the representations provided by the textual and visual branches. The resulting joint embedding space, learned through supervised training, is expected to capture the complexities of olfactory relationships in multimodal text and image data, particularly in cases where the relationship is positive despite the olfactory source not being explicitly depicted in the image.

Table 2 presents the number of parameters for each of the language-specific models used in our approach. This information is included to highlight the relative compactness of these models compared to the MM-LLMs that will be introduced later. Among the models, only the German branch differs in architecture as it is based on a distinct transformer variant, while the English, French, and Italian models are built upon BERT.

Table 2. Number of parameters for each language-specific model.

Model	EN	FR	DE	IT
Parameters	196 M	423 M	196 M	196 M

4.2. Mixture of Experts (MoE)

After training language-specific models, we explored whether combining their individual strengths could further enhance performance. Although all models were fine-tuned on the same multilingual dataset, each one is built upon a different pretrained text encoder, originally optimized for a specific language. These encoders differ in their pretraining sources and linguistic coverage, which leads to varied capabilities across languages. To take advantage of this diversity, we adopted a mixture-of-experts (MoE) approach. In this setup, a simple neural network, implemented as a linear layer, assigns dynamic weights to the outputs (scores) of the four language-specific models for each text–image pair. This allows the system to adaptively combine the predictions based on the input, effectively integrating the complementary strengths of the individual models.

To train the MoE network and to assign adequate weights to each input, cosine similarities are first computed for each text–image pair using the models obtained. Ideally, all of them should provide similar similarity scores since they have been trained on the same data, resulting in a consensual response. The final result of this network represents the probability that a text–image pair belongs to the positive class. Consequently, after obtaining the averaged result of four models, the final decision is made by simply applying a fixed classification threshold to this probability value so that, depending on whether it is above or below the threshold, a different category will be assigned. If it is above the threshold, it will be considered as a positive pair (matching smell) and as a negative pair (non-matching smell) otherwise, as shown in Figure 3.

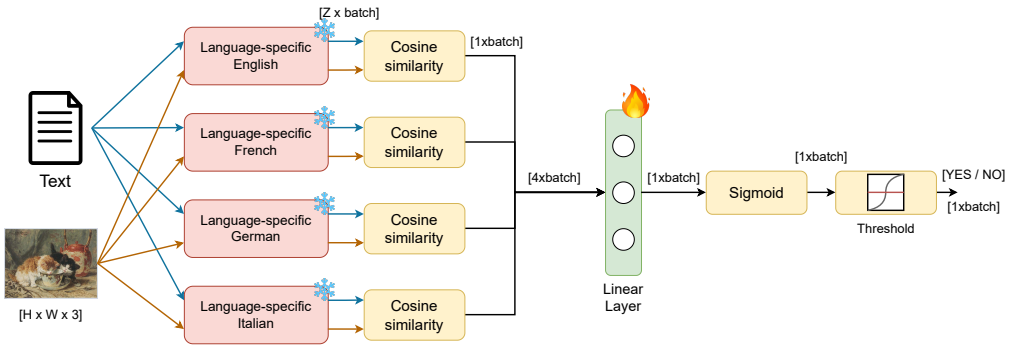


Figure 3. Architecture of the MoE model proposed with corresponding output dimensions. The orange lines correspond to image embeddings, and blue to text embeddings. Z represents the output of the projection layers of the language-specific models, so both representations belong to the same joint embedding space. After computing cosine similarities, a neural network is trained. The flame represents the part fine-tuned when trained and the snowflake those that are frozen.

In this context, as described in Section 4.1, it is necessary to determine the optimal threshold for each scenario. Although previously established thresholds produced satisfactory results for the language-specific models, this remains a potential source of concern

since it may indicate that the MoE network may partially inherit the biases acquired by the individual language models during training on imbalanced datasets.

4.3. MM-LLMs as Olfactory Experts

Encouraged by works such as those of Srinivasan et al. [34] and Kurfalı et al. [36] presented at MediaEval 2023, we explore the use of current state-of-the-art MM-LLMs to determine whether a text and an image contain references to the same smell sources. MM-LLMs can process both text and image data simultaneously, making them suitable for tasks that require understanding relationships between these modalities.

Figure 4 presents the architecture of the MM-LLM, highlighting its main components and providing an overview of how it processes and relates internal representations. A key distinction from language-specific models, and, by extension, MoEs, is that, rather than relying on global image–text embeddings, the MM-LLM operates on token-level interactions. This approach may offer an advantage in tasks that require identifying localized smell references across both modalities.

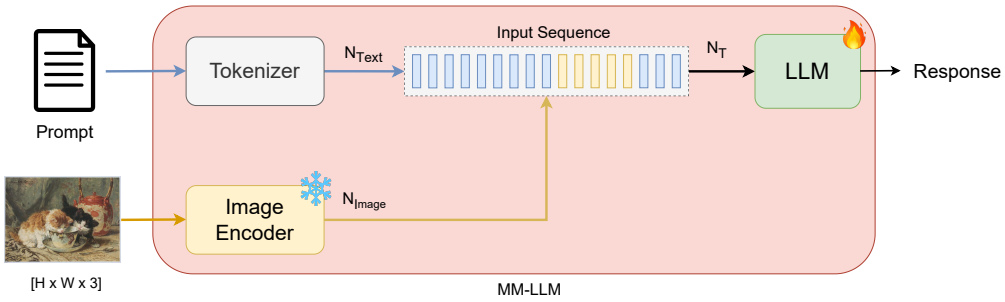


Figure 4. Architecture of the MM-LLM model proposed. The orange color refers to elements related to the visual part, and blue to the textual. In this case, both the text and the image are tokenized accordingly. Later, the sequence of tokens to be used as input for the LLM (of length $N_T = N_{Text} + N_{Image}$) is formed by concatenating both modality tokens. The flame represents the parts fine-tuned when trained and the snowflake those that are frozen.

4.3.1. Zero-Shot Evaluation

First of all, it is necessary to identify the most effective MM-LLM for our dataset, so we conducted a zero-shot evaluation by testing different prompts on the test data from the unbalanced negative class setup, which consists of 356 samples. The purpose of prompt engineering is to obtain a prompt that aligns the model capabilities obtained during its large-scale pretraining process well enough with the task under analysis. However, since we are working on a binary classification scenario where the models are required to answer with either "YES" or "NO", it is very unlikely that the output exactly matches one of the target labels. Consequently, the purpose of the zero-shot evaluation is to assess the model's performance without any fine-tuning on the specific task data for later adapting the model, demonstrating a higher understanding of the problem under analysis.

In particular, we tested two MM-LLMs, Kosmos-2 [42] and Qwen VL [13], in different configurations. The way these MM-LLMs process their input varies depending on how their pretraining process was carried out, so it is crucial to tailor the prompt format accordingly. This can be seen in Table 3, where the prompts used for the zero-shot evaluation of each model are presented, the majority derived from those proposed by Kurfal et al. [36].

Table 3. Prompts used for zero-shot evaluation on each model tested.

Model	Prompt
Kosmos-2	"<grounding>Determine if the following text and image share common elements, with a specific focus on smell sources. Look for entities such as objects, animals, fruits, or any other elements that could be potential sources of smells. Answer YES or NO. Text: <text>. Answer:"
Qwen-VL	"Question: Is there any object that appear in both the text and in Picture 1 with a specific focus on smell sources?. Text: <text>. Answer only YES or NO. ANSWER:"
Qwen-VL-Chat and Qwen-VL-Chat-Int4	"Question: Is there any object that appear in both the text and in <img>{im_path}</img>with a specific focus on smell sources?. Text: <text>. Answer only YES or NO. ANSWER:"

As commented previously, there is an important caveat regarding LLM output since it may not always strictly adhere to the expected format specified in the prompt, especially during zero-shot evaluation. In our case, while we aimed for binary "YES/NO" outputs, Qwen-VL sometimes generated full sentences explaining the text–image relationship, requested more information, or deviated from the task entirely. To address this, we implemented a simple post-processing step, validating only responses explicitly answering "YES/NO" in various forms. Anything else was considered incorrect. This highlights the potential limitations of zero-shot evaluation and the benefits of fine-tuning, as discussed in the next section.

4.3.2. Fine-Tuning Qwen-VL

After performing the zero-shot evaluation and considering the results obtained, we focused on fine-tuning Qwen-VL due to its pretrained superior performance. Its ability to embed the image directly within the prompt aligns well with our task, where referencing specific objects or entities in the image as potential smell sources is crucial. To fine-tune Qwen-VL, we employed the *SWIFT* [43] framework, which facilitates this process for Qwen-VL models using the LoRA method. In particular, LoRA is applied to the LLM while the vision encoder remains frozen. The primary advantage of LoRA is that, unlike traditional fine-tuning, it requires training a small number of parameters, which significantly reduces computational time and cost. This method is based on approximating the adaptation matrices with low-rank matrices. The influence of the newly adapted weights is controlled by two LoRA parameters: rank and alpha. The adaptation process can be defined as follows:

$$h = W_0 x + \frac{\alpha}{r} \Delta W x \tag{1}$$

In Equation (1), h represents the output of any layer with adaptation. W_0 is the original LLM weight matrix, and ΔW denotes the low-rank adaptation matrix that is learned during the LoRA process. The parameter r is the rank of the low-rank matrix, which controls the dimensionality of the adaptation, and α is a scaling factor that adjusts the influence of the adaptation. We have validated both rank and alpha to achieve the best possible performance in the challenge test data.

One key advantage of fine-tuning Qwen-VL is its ability to consistently provide the expected binary output ("YES/NO") for every text–image pair, eliminating the need for post-processing required during zero-shot evaluation (as discussed earlier). This observation suggests that even a single epoch of fine-tuning on our dataset led to significant convergence in terms of output format.

Since the MM-LLMs were primarily pretrained on English data (especially Qwen-VL [13], comprising 77.3% English and 22.7% Chinese textual content), all text in our dataset was translated to English for analysis. We achieved this by using a fine-tuned checkpoint

of mBART-large-50 (https://huggingface.co/facebook/mbart-large-50-one-to-many-mmt, accessed on 16 August 2025) [44].

5. Results and Discussion

The experiments were conducted on a NVIDIA RTX 4090 with 24GB VRAM. To evaluate our models, we followed a 5-fold cross-validation procedure to determine the best hyperparameters for our models. The macro F1-score, the primary metric used in the MUSTI challenge [8], was chosen for performance comparison in the in-house experiments, although more metrics are provided for the results evaluated on the official test dataset of the challenge.

In addition to the macro F1-score, we also employed Receiver Operating Characteristic (ROC) curves for a more comprehensive evaluation. This was particularly useful for assessing the performance of the language-specific models (Section 4.1) and the MoE approach (Section 4.2). ROC curves are particularly valuable in imbalanced class scenarios, which is the case for our dataset. To select the optimal threshold among the proposed approaches, we employed ROC curves implemented by scikit-learn (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html, accessed on 16 August 2025). The threshold that maximizes the geometric mean of sensitivity (Se, True Positive Rate) and specificity (Sp, True Negative Rate) can be defined as

$$\text{Threshold} = \arg\max_t \sqrt{\text{Se}(t) \times \text{Sp}(t)} \tag{2}$$

where $\text{Se}(t)$ and $\text{Sp}(t)$ represent the sensitivity and specificity at threshold t, respectively. The resulting optimal thresholds for each experimental setup are presented in Table 4, highlighting the induced biases present in the language-specific models trained on imbalanced datasets (unbalanced negative and unbalanced positive).

Table 4. Optimal thresholds for each specific setup.

Experiment Setup	Threshold
Unbalanced Negative	0.3
Balanced	0.5
Unbalanced Positive	0.6

To assess statistical significance, we compute confidence intervals (CIs) for the results obtained on the challenge test dataset. The macro F1-score is used as the primary evaluation metric, aligned with the official ranking criterion of MUSTI challenge. The method for calculating CIs is detailed in Equation (3). We consider two results to be statistically significantly different when their confidence intervals do not overlap.

$$\text{CI}(95\%) = \pm 1.96 \times \sqrt{\frac{\text{metric} \times (100 - \text{metric})}{N}} \tag{3}$$

5.1. Language-Specific Models

In order to evaluate our language-specific models, we considered two different experimental settings, which are described in Table 5. On the one hand, we employed an in-house dataset to evaluate our models before submitting them to be evaluated on the challenge test dataset. On the other hand, we used the official test dataset that provides a standardized way to compare with other publicly available results submitted to the challenge.

Table 5. Summary of experimental settings.

Experiment	Hyperparameters
5-fold Cross-Validation (CV) for our in-house experimental datasets (see Table 6)	Batch size: 16 Learning rate: 10^{-5} Early stopping: 10 epochs Early stop improvement: 0.005 Max epochs: 100
MUSTI Challenge official setup (see Tables 7–9)	Batch size: 16 Learning rate: 10^{-5} Early stopping: 10 epochs Early stop improvement: 0.005 Max epochs: 15

5.1.1. Results on In-House Datasets

Table 6 presents the macro F1-scores on our in-house dataset using the experimental settings and hyperparameters described in Table 5. In this case, it is the only metric evaluated as it is the one used to measure the performance of the submissions in the challenge.

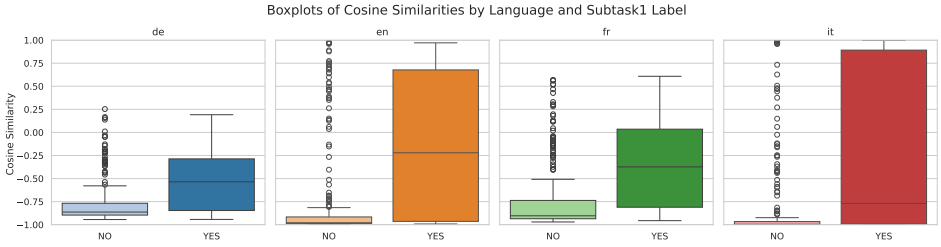
Table 6. Best average macro F1-scores obtained under 5-fold CV procedure for each model under our in-house datasets.

Exp. Setup	English	French	German	Italian	Average	MoE
Unbal. Neg.	0.6532	**0.7031**	0.6814	0.6311	0.6672	0.6226
Balanced	0.6889	**0.7253**	0.7015	0.6504	0.6915	0.6490
Unbal. Pos.	0.7043	0.6873	**0.7125**	0.6542	0.6895	0.5950

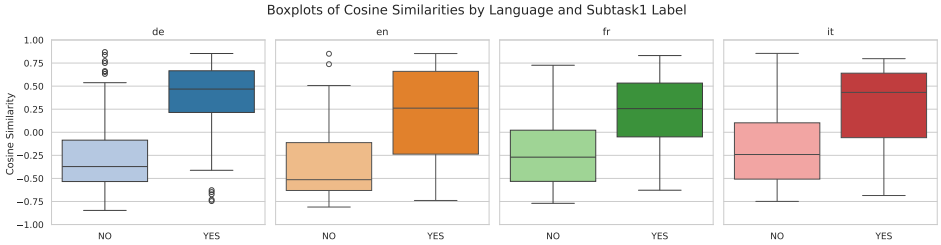
It can be observed that, in those cases where the model incorporates the **French** text encoder (CamemBERT), better performance is demonstrated across most setups (except for unbalanced positive), suggesting potential bias in the data or model architecture favoring French text (0.7253 is the best macro F1-score achieved in the balanced scenario), which can be explained due to its larger number of parameters alongside its larger internal representations (see Table 2).

Secondly, a clear trend emerged when analyzing performance across experimental setups. Balancing the data based on the target label (represented by the "Balanced" subset) consistently improved performance for all the language models, which is evident from their average performance (0.6915 average macro F1-score). This suggests that addressing class imbalance through data balancing techniques can be highly beneficial.

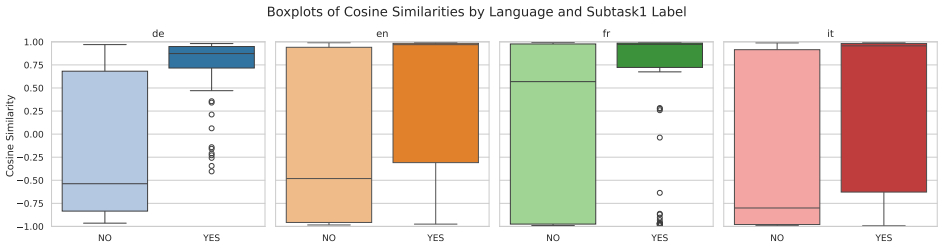
Regardless of the macro F1-score obtained by the models, we want to determine if the models are learning to represent elements that contain references to the same smell sources via providing aligned representations for both modalities. Figure 5 shows the distribution of cosine similarities produced by the best-performing model in each data distribution scenario for the best-performing language-specific model grouped by class label. Additionally, it provides valuable insights to understand the need for scenario-dependent thresholds to maximize performance. In Figure 5a, a clear shift in the cosine similarity distributions is observed, most notably in the German and French models, where the classes "YES" and "NO" exhibit relatively distinct distributions despite some degree of overlap. In Figure 5b, the separation between the two classes becomes more pronounced and symmetric for all languages, indicating a general improvement in the separability of the classes. In contrast, Figure 5c reveals lower separability, particularly in the English and Italian models, along with clear biases in the distributions.



(**a**) Unbalanced negative.



(**b**) Balanced.



(**c**) Unbalanced positive.

Figure 5. Cosine similarity distribution for every text–image pair on each test dataset for the experimental setups under analysis using best model obtained.

In general, these patterns suggest that the class imbalance significantly influences the internal representations of the model and can lead to biased decision boundaries. This supports the idea that majority-class downsampling may enhance performance, even at the cost of reducing the training set size.

Furthermore, when relating this visualization to the results in Table 6, it becomes evident that the balanced scenario, exhibiting the most consistent class separability across languages, is also the one in which the models achieve systematically better performance. This trend is further reflected in specific cases such as English, where the dataset biased towards the positive class ("matching smell") shows higher class separability and a correspondingly higher macro F1-score compared to the scenario biased towards the negative class (0.7043 vs. 0.6532).

5.1.2. Results on Challenge Dataset

The results evaluated on the challenge test dataset are reported in Tables 7–9, where the models have been trained on the entire in-house dataset for each setup and evaluated against the challenge test dataset. The results are consistent with the in-house validation, and additional metrics are provided for a more comprehensive analysis.

The **French** text encoder again performed better in unbalanced negative and balanced scenarios, achieving the best overall macro F1-score (**0.6401**) with the balanced setup. This could be due to CamemBERT's larger initial embedding size (1024 vs. 768 for others) and higher parameter count. Also, the German language-specific model showed consistent performance across the setups, suggesting robustness to data variations.

In the original unbalanced negative setup (see Table 7), the models favor the dominant class, with high performance for "NO" (average F1-score = **0.7220**) as expected, but struggles with the minority "YES" class (average F1-score = **0.4899**), resulting in an average macro F1-score of **0.6059**, indicating poor balance between classes, while the weighted F1-score (0.6495) and accuracy (0.6430) reflect better overall model performance, driven by the dominant class.

Table 7. Classification metrics for the unbalanced negative setup across languages on the official test split. Averages are computed excluding the MoE column.

Class	Metric	EN	FR	DE	IT	Avg.	MoE
NO	Precision	0.7589	0.7694	0.7880	0.7705	0.7717	0.6832
	Recall	0.7657	0.7818	0.5850	0.6064	0.6847	0.8640
	F1-score	0.7622	0.7755	0.6715	0.6787	0.7220	0.7630
YES	Precision	0.4739	0.5020	0.4171	0.4102	0.4508	0.2830
	Recall	0.4646	0.4843	0.6535	0.6024	0.5512	0.1180
	F1-score	0.4692	0.4930	0.5092	0.4880	0.4899	0.1667
Macro F1		0.6157	0.6342	0.5903	0.5834	0.6059	0.4648
Weighted F1		0.6707	0.6872	0.6208	0.6191	0.6495	0.5767
Accuracy		0.6716	0.6888	0.6064	0.6052	0.6430	0.6310

Across languages, English and French language-specific models tend to achieve higher scores on most metrics, while German and Italian show reduced performance likely to be underrepresented (see Figure 1b), suggesting language-specific challenges in capturing minority-class patterns in the skewed data.

When balancing the dataset (see Table 8), it can be seen how the performance of the models is better for the majority class overall since they achieve a higher average F1-score for the "NO" class (**0.7529**) in contrast to the average F1-score for "YES" (**0.4324**), which is lower, contrary to expectations.

Table 8. Classification metrics for balanced setup across languages on the official test split. Averages are computed excluding the MoE column.

Class	Metric	EN	FR	DE	IT	Avg.	MoE
NO	Precision	0.7288	0.7872	0.7784	0.7015	0.7490	0.6743
	Recall	0.8605	0.7281	0.7227	0.7317	0.7608	0.8479
	F1-score	0.7892	0.7565	0.7495	0.7163	0.7529	0.7512
YES	Precision	0.4902	0.4865	0.4728	0.3478	0.4493	0.2273
	Recall	0.2953	0.5669	0.5472	0.3150	0.4311	0.0984
	F1-score	0.3686	0.5232	0.5073	0.3306	0.4324	0.1374
Macro F1		0.5789	0.6401	0.6284	0.5234	0.5927	0.4443
Weighted F1		0.6578	0.6838	0.6739	0.5958	0.6528	0.5594
Accuracy		0.6839	0.6777	0.6679	0.6015	0.6578	0.6138

However, attending to the specific languages, we see that, in the case of French and German, performance is improved overall in all the metrics, suggesting that the process of removing examples has brought robustness to their internal representations. For the English and Italian models, performance drops significantly, especially in the case of Italian.

Regarding the unbalance towards the positive class scenario (see Table 9, the performance over the "YES" class is improved with respect to average F1-score (**0.4689**) and notably in recall (**0.6118**), suggesting the model is predicting more test samples as positive but without enough precision to claim that they improved their prediction capabilities.

This can be noticed in the global performance of the model, which significantly drops across global metrics such as average macro F1-score (**0.5352**), weighted F1-score (**0.5601**), and accuracy **(0.5483)**. While removing data for balancing can be effective, our results (particularly from the "unbalanced towards the positive" setup) show that excessive data removal can lead to a significant drop in performance for the positive class.

Table 9. Classification metrics for the unbalanced positive setup across languages on the official test split. Averages are computed excluding the MoE column.

Class	Metric	EN	FR	DE	IT	Avg.	MoE
NO	Precision	0.7235	0.7203	0.8218	0.7438	0.7524	0.6850
	Recall	0.5617	0.3363	0.5939	0.5403	0.5081	0.3345
	F1-score	0.6324	0.4585	0.6895	0.6259	0.6016	0.4495
YES	Precision	0.3536	0.3279	0.4450	0.3686	0.3738	0.3111
	Recall	0.5276	0.7126	0.7165	0.5906	0.6118	0.6614
	F1-score	0.4234	0.4491	0.5490	0.4539	0.4689	0.4232
Macro F1		0.5279	0.4538	0.6193	0.5399	0.5352	0.4363
Weighted F1		0.5671	0.4556	0.6456	0.5722	0.5601	0.4413
Accuracy		0.5510	0.4539	0.6322	0.5560	0.5483	0.4367

This highlights the need for potential data augmentation techniques, especially when aiming to improve positive class prediction capabilities while mitigating the negative effects of removing data from the majority class.

5.2. Mixture of Experts

We employed the same threshold selection procedure (Equation (2)) for the MoE approach, using the thresholds established for individual models (0.3, 0.5, and 0.6 for unbalanced negative, balanced, and unbalanced positive, respectively).

However, the MoE approach yielded lower performance compared to individual language-specific models on both the in-house dataset (Table 6) and the challenge dataset. This performance gap was even more pronounced with less training data.

These observations suggest that the current MoE architecture, which relies on a single dense layer to weight individual model outputs, lacks robustness and struggles to generalize, especially when data are limited.

This is especially remarkable when attending to the recall values in Tables 7–9, where the high values in both the unbalanced negative (**0.8640**) and balanced (**0.8479)** for the "NO" class and low values for the "YES" class (**0.1180** for unbalanced negative and **0.0984**) with low precision values suggest that the model presents a clear bias on predicting the majority class.

For the challenge dataset, the significant performance difference between MoE and individual models further highlights potential limitations in the MoE's ability to handle the task's complexity. Therefore, given the promising performance of individual language-specific models, we believe further analysis is required to find a better combination strategy that could benefit from the individual strengths of each model to provide more accurate predictions. For instance, adding a learnable threshold parameter that better aligns with the output of the linear layer that integrates cosine similarity values could be a promising yet simple and efficient way of enhancing performance. In fact, if shown to be beneficial, this adaptive thresholding mechanism could also be incorporated into the individual language-specific models to further enhance their effectiveness.

5.3. MM-LLMs

To evaluate the performance of the different MM-LLMs considered for the analysis, we conducted some assumptions to perform a fair comparison between them and the other approaches presented in this work, especially those related to zero-shot evaluation. As discussed in Section 4.3, the response provided is not necessarily limited to those specified in the given prompt. Although this problem appears only without fine-tuning, we have considered those cases with responses outside our binary target *"YES/NO"* as wrong predictions, labeling them as the opposite class of the ground-truth label. Furthermore, since the texts were translated into English for MM-LMs, the results are presented in Table 10 for both MM-LLMs and an English language-specific model trained with the original data.

Table 10. Results obtained for macro F1-score for MM-LLMs tested under different experimental setups following a zero-shot strategy.

MM-LLM Strategy	Exp. Setup	Kosmos-2	Qwen-VL	Qwen-VL-Chat-Int4	Qwen-VL-Chat
Zero-shot	Unbal. Neg.	0.3253	0.5581	0.4446	0.5661

5.3.1. Zero-Shot Performance on In-House Dataset

The zero-shot results (Table 10) show that all the Qwen-VL-based [13] models outperform Kosmos-2 [42]. This, combined with Qwen-VL's ability to directly embed images, led us to choose it for fine-tuning. Additionally, considering other metrics, apart from F1-score, it being used in the challenge, might be misleading due to the way we are evaluating the generated response. Consequently, the results reported in Table 10 must be considered not only as a measure of how well the model performs but also as a measure of response generation capability according to the instructions provided.

5.3.2. Fine-Tuning Qwen-VL-Chat on In-House Dataset

We fine-tuned and evaluated Qwen-VL-Chat and its quantized version in the same two different scenarios previously reported: (1) 5-fold CV for the in-house evaluation and (2) using all available data but the challenge official test split (see Table 5).

As expected, the base model outperforms the quantized version in the zero-shot setting (**0.5661** vs. **0.4446** in macro F1-score). However, after fine-tuning, both models achieve similar performance across setups, surpassing their pretrained states (see Table 10). Notably, the Qwen-VL-Chat model achieves a **0.8392** macro F1-score on the unbalanced negative class (all data setting). This highlights the potential benefits of fine-tuning quantized models for small datasets as they offer comparable performance with reduced computational resources. The results presented in Figures 6 and 7 have been obtained using a batch size of three and training for six epochs, demonstrating the average macro F1-score values for different configurations of LoRA parameters using Qwen-VL-Chat or Qwen-VL-Chat-Int4, respectively. The checkpoint used to obtain the results presented is the last one obtained through the training process.

Although LoRA Rank and LoRA Alpha are theoretically linked, with the importance given to newly learned weights being proportional to the ratio of LoRA Alpha to Rank (see Equation (1)), our experiments reveal a critical caveat. The performance gains associated with LoRA Alpha exhibit a limit determined by its absolute value. Contrary to the expected direct relationship, exceeding a certain Alpha value does not necessarily lead to further improvement. This might be because excessively high Alpha can overwhelm the base model, leading to instability and potentially detrimental effects. Therefore, finding the optimal balance between Rank and Alpha is crucial to maximize performance within this limitation.

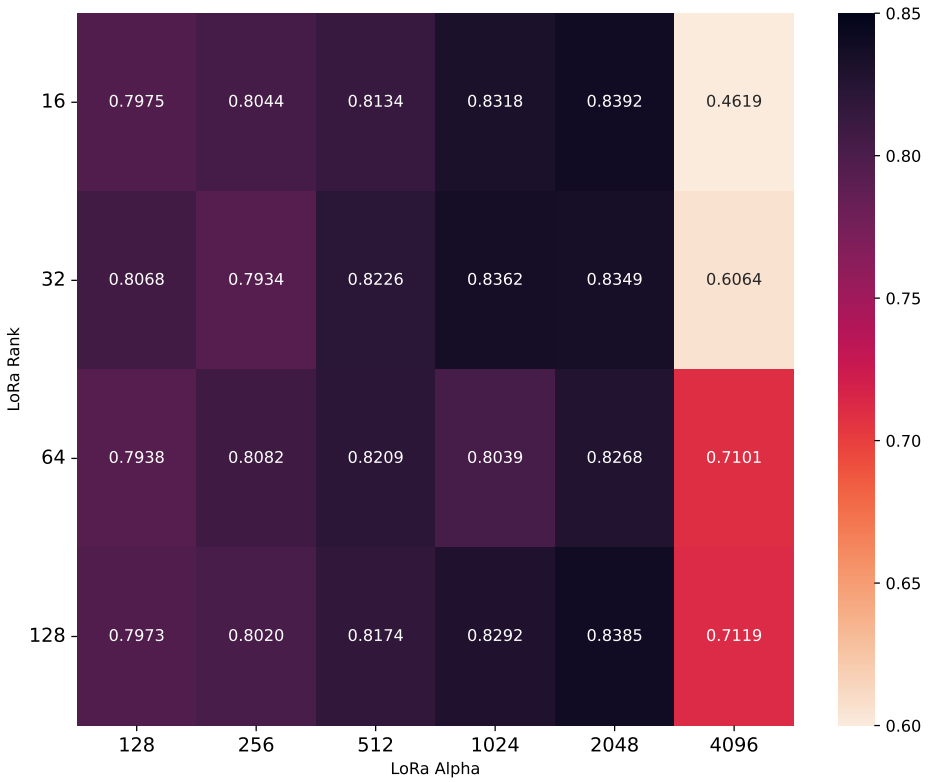


Figure 6. Average macro F1-scores obtained using 5-fold CV in our in-house experiments using Qwen-VL-Chat to obtain best LoRA parameters.

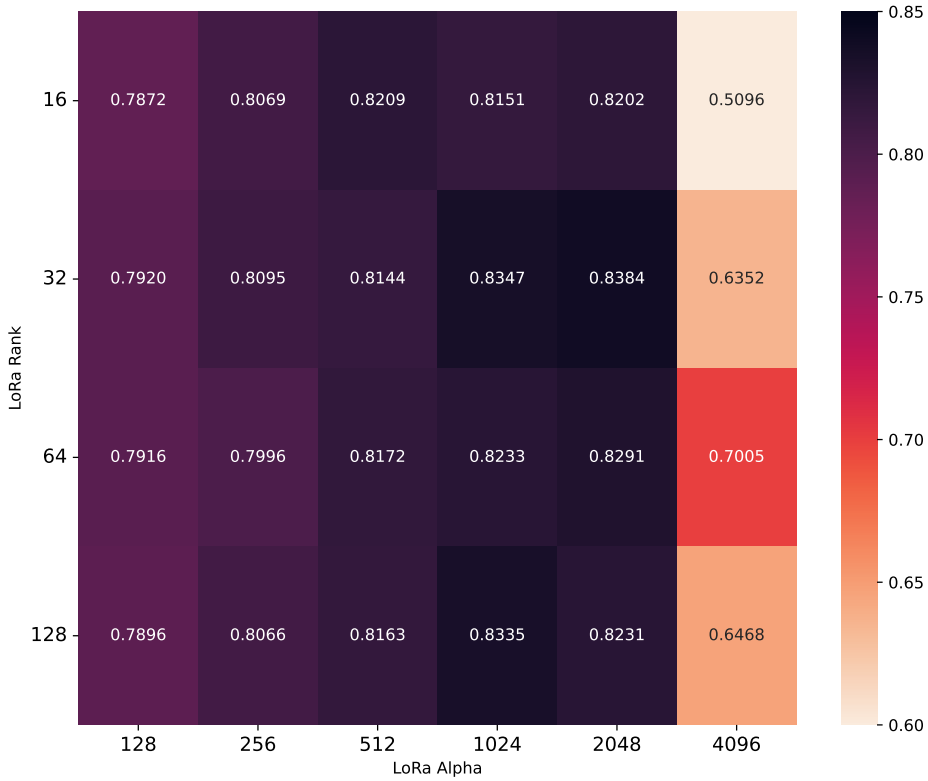


Figure 7. Average macro F1-scores obtained using 5-fold CV in our in-house experiments using Qwen-VL-Chat-Int4 to obtain best LoRA parameters.

5.3.3. Comparison with Language-Specific Models on In-House Dataset

Fine-tuning the multimodal large language model Qwen-VL-Chat resulted in the highest overall performance, achieving a macro F1-score of 0.8392 (as reported in Table 10). This significantly outperforms the best result obtained by any language-specific model in the balanced setup, with the French model reaching a macro F1-score of **0.7253**.

This also surpasses the top-performing English language-specific model (macro F1-score = **0.6936**), obtained after applying the corresponding English translation step in the in-house dataset. Notably, the translated language-specific model's performance is also better than the model without translation (0.6532 in Table 6) in the unbalanced towards negative scenario, suggesting that text translation can be beneficial for language-specific models in this task.

Despite being trained on the entire dataset, Qwen-VL-Chat exhibits superior performance, suggesting that its pretrained capabilities help to mitigate the influence of dataset biases during fine-tuning. It is also important to note that Qwen-VL-Chat has a substantially larger number of parameters (7B) compared to language-specific models (see Table 2), which likely contributes to a more nuanced and effective learning process.

5.3.4. Challenge Dataset Results

Finally, we evaluated our MM-LLMs in the challenge test dataset to compare with those reported in previous sections, where the best result achieved is **0.6401** using the French language-specific model regarding macro F1-score. The results are presented in Table 11, where both models under analysis outperform our previous best result, achieving **0.7464** macro F1-scores for the base model with LoRA Rank and Alpha of 16 and 2048, respectively, and, **0.7618** for the quantized version, 128 and 1024, respectively.

Table 11. Results obtained for macro F1-score when evaluating models on the challenge test dataset.

Class	Metric	Qwen-VL r = 128 α = 2048	Qwen-VL r = 16 α = 2048	Qwen-VL-Int4 r = 128 α = 1024	Qwen-VL-Int4 r = 32 α = 2048
NO	Precision	0.7622	0.8308	0.8325	0.8354
	Recall	0.8945	0.8694	0.8980	0.8623
	F1-score	0.8230	0.8497	0.8640	0.8486
YES	Precision	0.6242	0.6798	0.7286	0.6737
	Recall	0.3858	0.6102	0.6024	0.6260
	F1-score	0.4769	0.6432	0.6595	0.6490
Macro F1		0.6500	0.7464	0.7618	0.7488
Weighted F1		0.7149	0.7851	0.8001	0.7862
Accuracy		0.7355	0.7884	0.8057	0.7884

The runs submitted correspond with some of the best LoRA configurations, which yield the best results when performing 5-fold CV in our in-house experiments. It is also worth noting the improvements in precision observed when comparing the results to the scenario with data imbalance favoring the positive class, particularly for the language-specific models and the mixture-of-experts approach (see Table 9). In that setting, the recall increases, indicating that the models predicted the positive class ("YES") more frequently, although with low precision. In contrast, the MM-LLMs maintain higher precision while still effectively identifying positive instances, demonstrating greater robustness to the significant class imbalance without the need for data balancing techniques.

These findings underscore the effectiveness of MM-LLMs in olfactory understanding tasks and highlight the importance of fine-tuning such models for optimal performance, especially when compared to their base pretrained states. Moreover, the quantized model,

when fine-tuned, performs better than its base version, highlighting that, for this task, comparable performance can be obtained with a lighter model, resulting in decreased required computational resources and faster training processes. Consequently, Qwen-VL-Chat stands out as a promising candidate for further exploration in olfactory-related tasks, providing a foundation for further exploration of state-of-the-art MM-LLMs with improved capabilities.

When comparing our findings with those reported regarding the current state of the art in Kurfalı et al. [36], we did not find statistically significant differences (ours **0.7618** $\pm$ **0.0292** vs. Kurfalı et al. **0.7760** $\pm$ **0.0286**), indicating that our model closely competes with the best methodologies in tasks of olfactory understanding. Furthermore, our base model consists of Qwen-VL-Chat with 7 billion parameters, in contrast to the LLaVa [37] 13 billion they used. Specifically, our models can complete one training epoch in approximately 15 min on a NVIDIA RTX4090. This highlights the computational efficiency of our approach, enabling rapid experimentation.

6. Conclusions and Future Work

In this work, we presented innovative multimodal methods for determining whether a given text–image pair contains references to the same smell sources. Our first approach, which employs joint embedding spaces and language-specific models, demonstrated competitive performance across multiple languages. Notably, the French CamemBERT text encoder achieved a macro F1-score of **0.6401** on the MUSTI challenge test dataset with the balanced setup.

Additionally, we explored data balancing techniques and text translation as promising strategies for enhancing olfactory matching capabilities. These methods proved effective, particularly for the German and French models, which showed notable improvements when balancing the data. Despite these variations, our methods showed comparable performance overall when considering the average macro F1-scores across all the models.

To further optimize performance, we translated texts into English for subsequent training with an English language-specific model, which yielded a notable macro F1-score of **0.6936** in our in-house experiments under the Unbal Neg. setup. Therefore, translating demonstrates the potential of leveraging multilingual data to enhance model accuracy.

Secondly, we attempted to combine these models using a Mixture of Models approach. However, no improvement was observed through that combination. We attribute this to the limitation of a weighted linear combination of model outputs, suggesting that alternative combination methods, such as majority voting, routing weights based on language, or an adaptive thresholding mechanism, should be investigated to better utilize the individual strengths of the models.

In future research, conducting an ablation study on language-specific models could offer insights into the contribution of each branch to the task under investigation. Additionally, this would enable comparisons with other text and image encoders to determine which performs best individually. We could also explore the cultural variability of olfactory perception, considering how smell-related references may differ across languages and traditions. This could enhance the interpretability and applicability of multimodal systems in diverse cultural contexts.

Finally, the utilization of state-of-the-art MM-LLMs as olfactory experts has showcased promising results, particularly with a quantized version of Qwen-VL-Chat achieving a significant macro F1-score of **0.7618**. This approach demonstrates the potential of MM-LLMs to capture nuanced olfactory relationships and adapt responses to specific formats. Specifically, we compared their performance under two different scenarios: (1) first, by adopting a zero-shot evaluation scheme, and (2) by fine-tuning the models to the task at hand.

We observed that fine-tuning MM-LLMs on a downstream task can achieve comparable performance with reduced computational resources.

Regarding MM-LLMs, we consider it important to explore the adaptation of the vision encoder in the MM-LLM architecture to better align with the task at hand. We hypothesize that this adaptation could lead to significant improvements in performance given that the images in our dataset consist of paintings from the 17th to the 20th centuries. Most vision encoders have been trained primarily on datasets that contain real-world images, which may not fully capture the unique features and artistic styles present in historical artworks. By fine-tuning the vision encoder to recognize and interpret the nuances of these paintings, we can enhance the model's ability to associate smell relationships between text passages and the associated visual representation. Additionally, investigating alternative data balancing techniques, such as data augmentation and bagging, holds promise for enhancing model robustness, thereby enhancing model precision when predicting the minority class.

Addressing these challenges and pursuing these avenues of research will not only advance our understanding of olfactory representations in multimodal data but also pave the way for more effective scent-based information retrieval systems.

Author Contributions: Conceptualization, S.E.-R., M.G.-M. and F.F.-M.; methodology, S.E.-R., M.G.-M. and F.F.-M.; software, S.E.-R. and I.M.-F.; investigation, S.E.-R. and I.M.-F.; resources, M.G.-M. and F.F.-M.; writing—original draft preparation, S.E.-R.; writing—review and editing, S.E.-R., I.M.-F., M.G.-M., and F.F.-M.; supervision, M.G.-M. and F.F.-M.; project administration, M.G.-M. and F.F.-M.; funding acquisition, M.G.-M. and F.F.-M. All authors have read and agreed to the published version of the manuscript.

Funding: Sergio Esteban-Romero's research was supported by the Spanish Ministry of Education (FPI grant PRE2022-105516). The research of Iván Martín-Fernández was supported by the Universidad Politécnica de Madrid (Programa Propio I+D+i). This work was funded by Project ASTOUND (101071191—HORIZON-EIC-2021-PATHFINDERCHALLENGES-01) of the European Commission and by the Spanish Ministry of Science and Innovation through the projects GOMINOLA (PID2020-118112RB-C22), TRUSTBOOST (PID2023-150584OB-C21), and BeWord (PID2021-126061OB-C43), funded by MCIN/AEI/ 10.13039/501100011033 and by the European Union "NextGenerationEU/PRTR".

Data Availability Statement: The datasets presented in this article are not readily available because they have to be requested to the authors. Requests to access the datasets should be directed to ali.hurriyetoglu@gmail.com.

Acknowledgments: We would like to thank Ali Hürriyetoglu, the organizer of the MUSTI challenge, for his continuous support in evaluating runs, which was crucial for carrying out our research.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Lisena, P.; Schwabe, D.; van Erp, M.; Troncy, R.; Tullett, W.; Leemans, I.; Marx, L.; Ehrich, S.C. Capturing the Semantics of Smell: The Odeuropa Data Model for Olfactory Heritage Information. In *Proceedings of the Semantic Web*; Groth, P., Vidal, M.E., Suchanek, F., Szekley, P., Kapanipathi, P., Pesquita, C., Skaf-Molli, H., Tamper, M., Eds.; Springer: Cham, Switzerland, 2022; pp. 387–405.
2. Menini, S.; Paccosi, T.; Tonelli, S.; Van Erp, M.; Leemans, I.; Lisena, P.; Troncy, R.; Tullett, W.; Hürriyetoğlu, A.; Dijkstra, G.; et al. A multilingual benchmark to capture olfactory situations over time. In Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change, Dublin, Ireland, 26–27 May 2022; pp. 1–10.
3. Zinnen, M.; Madhu, P.; Kosti, R.; Bell, P.; Maier, A.; Christlein, V. Odor: The icpr2022 odeuropa challenge on olfactory object recognition. In Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; pp. 4989–4994.

4. Wang, P.Y.; Sun, Y.; Axel, R.; Abbott, L.; Yang, G.R. Evolving the olfactory system with machine learning. *Neuron* **2021**, *109*, 3879–3892. [CrossRef]
5. 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**, *381*, 999–1006. [CrossRef] [PubMed]
6. Ameta, D.; Kumar, S.; Mishra, R.; Behera, L.; Chakraborty, A.; Sandhan, T. Odor classification: Exploring feature performance and imbalanced data learning techniques. *PLoS ONE* **2025**, *20*, e0322514. [CrossRef]
7. Tan, H.; Zhou, Y.; Tao, Q.; Rosen, J.; van Dijken, S. Bioinspired multisensory neural network with crossmodal integration and recognition. *Nat. Commun.* **2021**, *12*, 1120. [CrossRef] [PubMed]
8. Hürriyetoglu, A.; Paccosi, T.; Menini, S.; Zinnen, M.; Lisena, P.; Akdemir, K.; Troncy, R.; van Erp, M. MUSTI—Multimodal Understanding of Smells in Texts and Images at MediaEval 2022. In *Working Notes Proceedings of the MediaEval 2022 Workshop, Bergen, Norway and Online, 12–13 January 2023*; Hicks, S., de Herrera, A.G.S., Langguth, J., Lommatzsch, A., Andreadis, S., Dao, M., Martin, P., Hürriyetoglu, A., Thambawita, V., Nordmo, T.S., et al., Eds.; 2022; Volume 3583, CEUR Workshop Proceedings. Available online: https://ceur-ws.org/Vol-3583/ (accessed on 16 August 2025).
9. Masaoka, Y.; Sugiyama, H.; Yoshida, M.; Yoshikawa, A.; Honma, M.; Koiwa, N.; Kamijo, S.; Watanabe, K.; Kubota, S.; Iizuka, N.; et al. Odors Associated With Autobiographical Memory Induce Visual Imagination of Emotional Scenes as Well as Orbitofrontal-Fusiform Activation. *Front. Neurosci.* **2021**, *15*, 709050. [CrossRef]
10. Ehrich, S.; Verbeek, C.; Zinnen, M.; Marx, L.; Bembibre, C.; Leemans, I. Nose-First. Towards an Olfactory Gaze for Digital Art History. In *LDK 2021 LDK Workshops and Tutorials, Zaragoza, Spain, 1–4 September 2021*; Carvalho, S., Rocha Souza, R., Eds. 2021; Volume 3064, CEUR Workshop Proceedings. Available online: https://ceur-ws.org/Vol-3064/ (accessed on 16 August 2025).
11. Esteban-Romero, S.; Martín-Fernández, I.; Bellver-Soler, J.; Gil-Martín, M.; Martínez, F.F. Multimodal and Multilingual Olfactory Matching based on Contrastive Learning. In Proceedings of the MediaEval, Amsterdam, The Netherlands and Online, 1–2 February 2024.
12. 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. *arXiv* **2021**, arXiv:2103.00020. [CrossRef]
13. Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; Zhou, J. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv* **2023**, arXiv:2308.12966.
14. Desai, K.; Johnson, J. VirTex: Learning Visual Representations from Textual Annotations. *arXiv* **2021**, arXiv:2006.06666. [CrossRef]
15. Zhang, Y.; Jiang, H.; Miura, Y.; Manning, C.D.; Langlotz, C.P. Contrastive Learning of Medical Visual Representations from Paired Images and Text. *arXiv* **2022**, arXiv:2010.00747. [CrossRef]
16. Hürriyetoglu, A.; Novalija, I.; Zinnen, M.; Christlein, V.; Lisena, P.; Menini, S.; van Erp, M.; Troncy, R. The MUSTI Challenge @ MediaEval 2023—Multimodal Understanding of Smells in Texts and Images with Zero-shot Evaluation. In Proceedings of the MediaEval 2023 Workshop, Amsterdam, The Netherlands and Online, 1–2 February 2024.
17. Akdemir, K.; Hürriyetoglu, A.; Troncy, R.; Paccosi, T.; Menini, S.; Zinnen, M.; Christlein, V. Multimodal and Multilingual Understanding of Smells using VilBERT and mUNITER. In *Working Notes, Proceedings of the MediaEval 2022 Workshop, Bergen, Norway and Online, 12–13 January 2023*; Hicks, S., de Herrera, A.G.S., Langguth, J., Lommatzsch, A., Andreadis, S., Dao, M., Martin, P., Hürriyetoglu, A., Thambawita, V., Nordmo, T.S., et al., Eds. 2022; Volume 3583, CEUR Workshop Proceedings. Available online: https://ceur-ws.org/Vol-3583/ (accessed on 16 August 2025).
18. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *arXiv* **2019**, arXiv:1908.02265.
19. Liu, F.; Bugliarello, E.; Ponti, E.; Reddy, S.; Collier, N.; Elliott, D. Visually Grounded Reasoning across Languages and Cultures. In Proceedings of the NeurIPS 2021 Workshop on ImageNet: Past, Present, and Future, Online, 13–14 December 2021.
20. Xie, N.; Lai, F.; Doran, D.; Kadav, A. Visual entailment task for visually-grounded language learning. *arXiv* **2018**, arXiv:1811.10582.
21. Mirunalini, P.; Sanjhay, V.; Rohitram, S.; Rohith, M. MUSTI-Multimodal Understanding of Smells in Texts and Images Using CLIP. In Proceedings of the MediaEval, Amsterdam, The Netherlands and Online, 1–2 February 2024.
22. Ngoc-Duc, L.; Minh-Hung, L.; Quang-Vinh, D. Handle the problem of ample label space by using the Image-guided Feature Extractor on the MUSTI dataset. In Proceedings of the MediaEval, Amsterdam, The Netherlands and Online, 1–2 February 2024.
23. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
24. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
25. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1 (long and short papers), pp. 4171–4186.
26. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165. [CrossRef]

27. Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. The llama 3 herd of models. *arXiv* **2024**, arXiv:2407.21783. [CrossRef]
28. Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M.S.; Love, J.; et al. Gemma: Open models based on gemini research and technology. *arXiv* **2024**, arXiv:2403.08295. [CrossRef]
29. Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. Qwen Technical Report. *arXiv* **2023**, arXiv:2309.16609. [CrossRef]
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2023**, arXiv:1706.03762. [PubMed]
31. Chu, Y.; Xu, J.; Yang, Q.; Wei, H.; Wei, X.; Guo, Z.; Leng, Y.; Lv, Y.; He, J.; Lin, J.; et al. Qwen2-Audio Technical Report. *arXiv* **2024**, arXiv:2407.10759. [CrossRef]
32. Esteban-Romero, S.; Martín-Fernández, I.; Gil-Martín, M.; Griol-Barres, D.; Callejas-Carrión, Z.; Fernández-Martínez, F. LLM-Driven Multimodal Fusion for Human Perception Analysis. In Proceedings of the 5th on Multimodal Sentiment Analysis Challenge and Workshop: Social Perception and Humor, Melbourne, VIC, Australia, 28 October 2024; MuSe'24, pp. 45–51. [CrossRef]
33. Martín-Fernández, I.; Esteban-Romero, S.; Bellver-Soler, J.; Fernández-Martínez, F.; Gil-Martín, M. Larger Encoders, Smaller Regressors: Exploring Label Dimensionality Reduction and Multimodal Large Language Models as Feature Extractors for Predicting Social Perception. In Proceedings of the 5th on Multimodal Sentiment Analysis Challenge and Workshop: Social Perception and Humor, Melbourne, VIC, Australia, 28 October 2024; pp. 20–27.
34. rinivasan, D.; Subhashree, M.; Mirunalini , P.; Jaisakthi, S.M. Multimodal Learning for Image-Text Matching: A Blip-Based Approach. In Proceedings of the MediaEval, Amsterdam, The Netherlands and Online, 1–2 February 2024.
35. Li, J.; Li, D.; Xiong, C.; Hoi, S. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *arXiv* **2022**, arXiv:2201.12086.
36. Kurfalı, M.; Olofsson, J.K.; Hörberg, T. Enhancing Multimodal Language Models with Olfactory Information. In Proceedings of the MediaEval, Amsterdam, The Netherlands and Online, 1–2 February 2024.
37. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual Instruction Tuning. *arXiv* **2023**, arXiv:2304.08485.
38. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**, arXiv:2106.09685.
39. Ridnik, T.; Ben-Baruch, E.; Noy, A.; Zelnik-Manor, L. ImageNet-21K Pretraining for the Masses. *arXiv* **2021**, arXiv:2104.10972.
40. Song, K.; Tan, X.; Qin, T.; Lu, J.; Liu, T.Y. MPNet: Masked and Permuted Pre-training for Language Understanding. *arXiv* **2020**, arXiv:2004.09297. [CrossRef]
41. Martin, L.; Muller, B.; Suárez, P.J.O.; Dupont, Y.; Romary, L.; de la Clergerie, É.V.; Seddah, D.; Sagot, B. CamemBERT: A Tasty French Language Model. *arXiv* **2019**, arXiv:1911.03894.
42. Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; Wei, F. Kosmos-2: Grounding Multimodal Large Language Models to the World. *arXiv* **2023**, arXiv:2306.14824. [CrossRef]
43. Team, T.M. SWIFT:Scalable lightWeight Infrastructure for Fine-Tuning. 2024. Available online: https://github.com/modelscope/swift (accessed on 16 August 2025).
44. Tang, Y.; Tran, C.; Li, X.; Chen, P.J.; Goyal, N.; Chaudhary, V.; Gu, J.; Fan, A. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning. *arXiv* **2020**, arXiv:2008.00401. [CrossRef]