How LLMs See People

Carlos Roxo[1,2], João Marcos[1,3], and Nuno Gonçalves[1,4]

[1] Institute of Systems and Robotics, University of Coimbra, Portugal, 3030-290
[2] carlos.roxo@isr.uc.pt
[3] joao.marcos@isr.uc.pt
[4] nunogon@deec.uc.pt

Abstract. This study investigates the linguistic patterns in human subject descriptions that came out from large language models with vision capabilities. Descriptions generated by two vision-capable LLMs, Qwen2.5-VL-72B-Instruct and Llama3.2-vision:11b, were examined for a subset of the CelebA dataset. Using word frequency and clustering analyses, we have identified distinct common topics in descriptions of people, including facial features, hair characteristics, clothing, body structure, posture, and contextual environment. Our findings showed differences in how these models organize descriptive concepts, with Llama3.2 demonstrating more gender-centric descriptions compared to Qwen2.5's focus on objective physical attributes. These patterns may reveal underlying conceptual frameworks that shape how LLMs represent human subjects. Our analysis contributes to understanding representation in multimodal AI systems and has implications for reducing bias in descriptions of people.

Keywords: large language models · vision models · person description

1 Introduction

Large Language Models (LLMs) have evolved significantly in recent years, using transformer architectures and massive datasets to achieve remarkable performance in natural language processing (NLP). Models such as GPT-4 (OpenAI) [12], LLaMA (Meta) [13], and PaLM (Google) [14] demonstrate high proficiency in text generation, reasoning, and multimodal tasks. A key advancement is their ability to process and generate human-like text based on contextual understanding, benefiting from self-attention mechanisms [2]. Recent research explores the integration of LLMs with vision models for multimodal understanding. Models like DALL·E [15] and CLIP [16] show that neural networks can bridge textual and visual modalities, enabling tasks such as text-to-image generation and zero-shot image recognition.

One particularly intriguing application is the generation of descriptions that may encapsulate aspects of human appearance, behavior, and identity [4]. The increasing concerns about privacy, security and bias issues of Facial Recognition Systems (FRS) [21] led the computer vision (CV) and artificial inteligence (AI) community to develop broader and more robust methods for the automatic analysis of people. The way LLMs process this task can reveal a lot on their internal representation of human categories and the embedded potential biases. In this study, we examine descriptions generated by two multimodal models, Qwen2.5-VL-72B-Instruct [5] and Llama3.2-vision:11b [6], for a subset of images from the CelebA dataset [7]. Through the analysis of word frequencies, co-occurrence patterns, and semantic clustering, we identify and quantify the primary topics of interest that may signify how these models represent human subjects.

Our research addresses several key questions:

1. What are the primary descriptive categories that LLM models use when characterizing people in images?
2. How do LLM models organize and prioritize different aspects of human appearance?
3. Are there significant differences in how different models describe people?
4. To what extent do demographics, like ethnicity or gender, influence descriptive patterns?

By addressing these questions, our study aims to contribute to a deeper understanding of how LLM based systems represent people, an essential consideration as these technologies become increasingly integrated into applications that interact with and make decisions about human beings [8]. Identifying patterns in person descriptions may help detect potential biases, inform more balanced model development, and guide responsible deployment of vision-language systems. Bias analyses have direct implications for critical domains where bias mitigation is essential—including law enforcement facial recognition systems, healthcare diagnostics, insurance risk assessment, and hiring processes. Furthermore, such research aligns with emerging regulatory frameworks such as the EU AI Act [21], which specifically targets high-risk AI applications and requires transparency, fairness, and accountability in systems that make critical decisions on individuals.

This paper is organized as follows. Section 2 provides background information on facial description research and bias detection in vision-language models. Section 3 details our data gathering methodology, including dataset selection, model choices, and data collection process. Section 4 explains our analytical approach to examine descriptive patterns. Sections 5 and 6 present our analysis results for Qwen2.5-VL-72B-Instruct and Llama3.2-vision:11b, examining how these models organize concepts when describing people. Finally, Section 7 discusses our findings, examines implications for multimodal AI systems, and suggests directions for future research.

2 Background

Research on facial description and perception provides important context for understanding how AI systems characterize people. Tyler et al. (2022) [9] recently investigated the efficacy of verbal face descriptions in identification tests.

The authors discovered that, although individuals were able to produce reliable descriptions, their ability to recognize faces from these descriptions was only moderately accurate. This may imply intrinsic challenges in accurately converting visual facial information into verbal descriptions. Building on this understanding of feature importance, Diego-Mas et al. (2020) [10] investigated the impact of individual facial features on perception. Their findings indicate that features such as the eyes and mouth significantly influence judgments about traits like attractiveness and emotional expression. This selective attention to specific features resembles potential patterns in how AI systems might prioritize certain facial characteristics when generating descriptions. Additionally, Stolier et al. (2018) [11] demonstrated that conceptual associations between personality traits affect how we perceive faces. The study suggests that our beliefs about trait correlations influence the visual features we attend to when forming impressions, highlighting a dynamic interplay between conceptual knowledge and visual perception.

On identifying and mitigating biases in image captioning systems, Zhao et al. (2021) [22] conducted a comprehensive analysis of racial biases in image captioning, demonstrating how these models often perpetuate stereotypes and generate different descriptions based on perceived race. Their work emphasizes the need for more inclusive datasets and more nuanced evaluation metrics that can detect subtle forms of bias. Hamidieh et al. (2024) [23] categorized social biases in vision-language models across numerous domains, including appearance, behavior, education, wealth, criminal justice, healthcare, media portrayal, and occupation, demonstrating the pervasive nature of these biases throughout different applications.

The bidirectional relationship between text and image is also relevant to our work. While our study examines how models generate text from facial images, related research explores the inverse process. Sharma (2024) [17] developed attention-based GAN models that achieve impressive inception scores and FID metrics for text-to-image generation. Sádaba-Campo and Gómez-Moreno (2025) [18] explored the use of generative neural networks, including both GANs and diffusion models, to automate composite sketch creation with tools like DALL-E and Stable Diffusion. A VICE investigation (2023) [19] highlighted significant risks of bias and misidentification in AI-generated forensic sketches, particularly in systems utilizing OpenAI's DALL-E 2. These concerns underline the importance of understanding how multimodal systems conceptualize and describe human subjects.

Understanding both directions of this text-image relationship is crucial, as biases can be amplified through iterative processes where biased descriptions generate biased images, which in turn reinforce biased descriptions. This cyclical reinforcement of bias highlights the importance of interventions at multiple points in vision-language systems.

3 Data Collection

3.1 Dataset

This study used the CelebA (Large-scale CelebFaces Attributes) dataset [7], which contains over 200,000 celebrity face images annotated with 40 attribute labels. To avoid systematic bias in the selection process and ensure a broad representation of different individuals, the images used were randomly selected from the dataset.

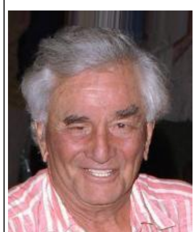
Example Image	Qwen2.5-VL-72B-Instruct's description
	The individual in the image appears to be an older adult with short, light-colored hair that is styled neatly. The visible part of their face shows a broad smile, revealing teeth, and they have prominent laugh lines around their mouth and eyes, suggesting frequent smiling or laughter. Their skin tone is fair. In terms of body type, only the upper torso is visible. The person seems to have a medium build, as indicated by the fit of their clothing. They are wearing a short-sleeved, collared shirt with a pink and white striped pattern. The shirt appears to be made of a lightweight fabric, suitable for casual or semi-formal occasions. Their posture is upright, and they seem to be standing confidently. There are no visible accessories such as jewelry, watches, or glasses on the parts of the body that are shown. The background is blurred, but it appears to be an indoor setting with other people present, possibly at a social gathering or event.
Example Image	Llama3.2-vision:11b's description
	The image presents a close-up portrait of a woman with long brown hair that cascades down her shoulders. The lighting on the right side of the photo accentuates her features. **Facial Features:** * Her eyes are cast downward, gazing slightly to the left. * Her eyebrows are neatly trimmed and complemented by light makeup. * A subtle smile plays on her lips, hinting at a sense of contentment or amusement. **Attire:** * The subject wears an intricately patterned blouse featuring a vibrant mix of colors, including yellow, red, blue, green, white, and black. * Her attire is not fully visible in the frame but appears to be a formal or semi-formal outfit. **Overall Impression:** The woman's demeanor exudes confidence and poise. The blurred background behind her creates a sense of intimacy and focus on her individuality.

Table 1: Example of two facial description generated by Qwen2.5-VL-72B-Instruct and Llama3.2-vision:11b, respectively

3.2 Model Selection

Two vision-capable large language models were selected for this analysis:

– Qwen2.5-VL-72B-Instruct [5]: A 72 billion parameter multimodal model developed by Alibaba Cloud, capable of processing both text and images.
– Llama3.2-vision:11b [6]: Meta AI's 11 billion parameter multimodal model, which integrates vision capabilities with Llama's language generation abilities.

We have selected the aformentioned models for this analysis based on several key considerations:

1. Architectural Diversity: these models represent different approaches to multimodal integration. Qwen2.5 employs a unified transformer architecture with interleaved processing of visual and textual tokens, while Llama3.2 adopts a modular approach with separate visual encoders and language decoders connected through projection layers.
2. Parameter Scale: by including both a larger model (Qwen2.5) and a smaller model (Llama3.2), we can investigate whether model size influences descriptive patterns.
3. Model Accessibility: Llama3.2 is small enough to run locally, while Qwen2.5's web interface allows for unlimited prompts with images, facilitating the gathering of large amounts of image-caption pairs.

Although there are more vision-capable LLMs, the selected models provide a useful comparative framework.

3.3 Data Collection Process

To maintain consistency across all descriptions, we used a standardized, zero-shot prompt for both models: "**Describe this person in the most detailed and objective way possible based on the visible parts of their body. Cover facial features, body type (if applicable), posture, clothing, accessories, and any other observable details, without making subjective judgments or assumptions beyond what is visible.**". This prompt was chosen to obtain objective descriptions. The same prompt was used for all images to ensure that variations in the descriptions would be attributable to differences in the models' processing. 300 descriptions were collected from Qwen2.5 and 1000 descriptions from Llama3.2, each of a different person. Each image was processed independently, with no context from previous descriptions or images.

The difference in the number of descriptions collected from each model was due to practical constraints. Llama3.2, being a smaller model, could be run locally, allowing for automated large-scale data collection, which led us to gather 1000 descriptions. In contrast, Qwen2.5 is a significantly larger model that was more challenging to run locally, requiring manual collection of its descriptions using Qwen's web interface and Hugging Face's API, limiting the dataset to 300 entries. To assess the impact of this discrepancy, we compared an analysis using only 300 descriptions from Llama3.2 to the 1000 description set. The results indicated very little discrepancy, with the main difference being that, with 1000 descriptions, themes became more discernible. However, Qwen2.5's descriptions already exhibited well-discernible themes with just 300 descriptions, so we believe that 300 descriptions provide a valid basis for comparison.

4 Methodology

This study employed two main analytical approaches to examine how vision-capable LLMs describe people: a word frequency analysis and a clustering analysis. These methods allowed us to identify patterns in descriptive language and reveal the underlying conceptual frameworks used by the models.

4.1 Word Frequency Analysis

To identify the most common descriptive terms used by each model, we calculated word frequencies across all descriptions. The process involved:

1. Tokenizing each description into individual words
2. Converting all words to lowercase to ensure consistency
3. Removing common stop words (e.g., "the," "a," "is") that do not contribute with meaningful content
4. Counting occurrences of each remaining word
5. Ranking words by frequency to identify the most common descriptive terms

This analysis provided initial insights into which physical attributes and characteristics each model prioritized when describing human subjects [25].

4.2 Clustering Analysis

To understand how models organize concepts when describing people, we performed a clustering analysis based on word co-occurrence patterns. We employed two different approaches to measure co-occurrence, adapting to the structural differences in descriptions generated by each model. Similar clustering approaches have been used to extract major themes from textual data [26]

Sentence-Based Co-occurrence For the descriptions generated by Qwen2.5, we applied a sentence-based co-occurrence measure. For instance, the word *upper* appears in approximately 75% of the phrases that contain *body*, whereas *body* appears in only 37% of the phrases that contain *upper*. Evidently, while *body* can be preceded by various terms, *upper* is strongly associated with just the word *body*. To account for this, the sentence-based co-occurrence score is often determined by selecting the maximum value between these two proportions. The computation of this score follows Equation 1.

$$\text{Co-occurrence} = \frac{\text{Phrases with both words}}{\min(\text{Phrases with word one}, \text{Phrases with word two})} \quad (1)$$

This approach effectively captured relationships between words that frequently appeared together within the same sentence.

Sliding Window Co-occurrence For Llama3.2's descriptions, a sliding window approach with a window size of 5 words proved more effective for assessing word co-occurrence than the sentence-based approach used for Qwen2.5. This methodological adaptation was necessary due to structural differences in the descriptions generated by each model. Initial application of the sentence-based co-occurrence method to Llama3.2's outputs resulted in poorly differentiated clusters with excessive conceptual overlap, making interpretation difficult. This difference in effective methodologies may reflect the distinct writing styles of each model. Llama3.2 tended to generate more bullet-point-style descriptions with shorter sentences, where related concepts were often separated in several sentences but remained within proximity. In contrast, Qwen2.5 produced longer descriptions with clearer sentence boundaries where related concepts were typically contained within the same sentence. The window-based approach for Llama3.2 allowed us to capture these proximity relationships more effectively than sentence boundaries would permit. The computation of this score follows Equation 2.

$$\text{Co-occurrence} = \frac{\text{Windows with both words}}{\min(\text{Windows with word 1}, \text{Windows with word 2})} \tag{2}$$

Community Detection For both co-occurrence measurement approaches, we applied the Louvain community detection method [27] to identify clusters of related words. The Louvain method optimizes the modularity score, a scalar value between -1 and 1 that measures the density of links within communities compared to the density of links between communities, defined in equation 3, where A_{ij} represents the weight of the edge between vertices i and j, $k_i = \sum_j A_{ij}$ is the sum of the weights of the edges attached to vertex i, c_i is the community to which vertex i is assigned, the δ-function is defined as $\delta(u, v) = 1$ if $u = v$ and 0 otherwise, $m = \frac{1}{2} \sum_{i,j} A_{ij}$ and γ is the resolution parameter [27,29]. By default, modularity is computed with γ=1, but during optimization, this parameter can be adjusted to achieve community partitions of different sizes. A higher modularity score indicates a better division into communities. In our experiments, the resolution parameter was set to 1.0 for Qwen2.5 and 0.9 for Llama3.2. The following steps were followed:

1. Select the 100 most frequent words after stop word removal
2. Create a graph where nodes represent words and edges represent co-occurrence scores
3. Apply the Louvain algorithm with appropriate resolution parameters
4. Evaluate clusters using modularity scores

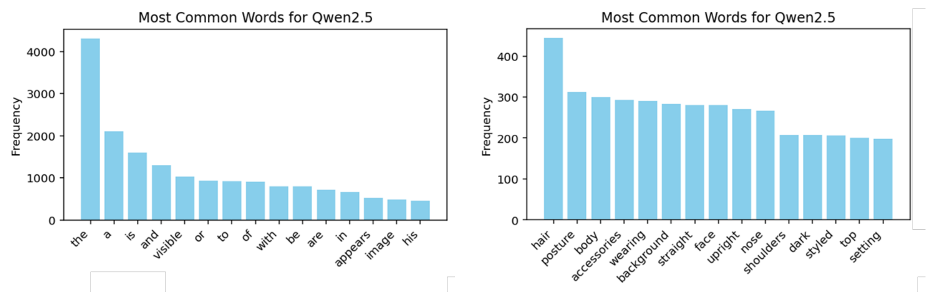
$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \gamma \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{3}$$

The resulting clusters revealed distinct semantic categories and conceptual frameworks employed by each model when describing human subjects.

5 Qwen2.5-VL-72B-Instruct Analysis Results

5.1 Word Frequency Analysis

After removing stop words, the most frequent descriptive terms in Qwen2.5's descriptions were analyzed. As shown in Figure 1, *hair* emerged as the most frequent descriptive term, followed by *posture*, *body*, *accessories*, and *wearing*.



(a) Word frequency distribution before stop word filtering

(b) Word frequency distribution after stop word filtering

Fig. 1: Word frequency distributions before and after stop word filtering Qwen2.5's descriptions

5.2 Clustering Analysis

Using the sentence-based co-occurrence method and Louvain community detection algorithm, we identified seven distinct clusters with a modularity score of 0.480, illustrated in figure 2. The words in each cluster, ordered by decreasing frequency of occurrence, are the following:

– **Cluster 1:** hair, styled, skin, light, tone, appear, short, brown, neatly, adult, male, side, parted, slight.
– **Cluster 2:** posture, upright, shoulders, head, relaxed, standing, forward, stance, facing, held, confident.
– **Cluster 3:** body, type, upper, torso, slender, build, average.
– **Cluster 4:** accessories, jewelry, watches, frame, within, glasses, adornments.
– **Cluster 5:** wearing, dark, top, shirt, white, jacket, clothing, garment, collar, black, color, neckline, fabric, portion, material, smooth, neck, made, dress, design, thin, attire, sleeveless, around, chest, blue, subtle, pattern, small, earrings.
– **Cluster 6:** background, setting, lighting, blurred, context, event, formal, details, indoor, text, might, even, back, focus, photo.
– **Cluster 7:** straight, face, nose, eyes, lips, features, full, facial, partially, eyebrows, neutral, arched, closed, expression, mouth, looking

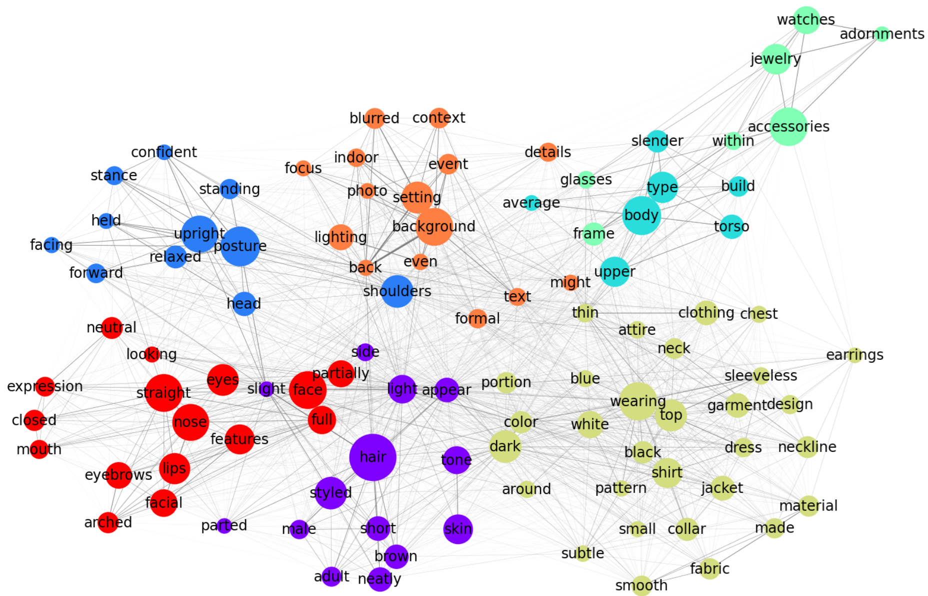


Fig. 2: Graphical representation of Qwen2.5's word communities, visualized by the Fruchterman-Reingold algorithm[28]

5.3 Findings and Implications

The clustering results suggest that Qwen2.5 organizes person descriptions into distinct semantic categories, prioritizing observable physical features over subjective interpretations. These categories can be interpreted as follows:

1. **Cluster 1: Hair & Style -** Describes characteristics related to hair and style (*styled*, *short*, *neatly*), as well as complexion (*skin*, *tone*).
2. **Cluster 2: Posture** – Focuses on how the body is positioned and oriented, with terms describing posture and direction (*posture*, *upright*, *standing*, *forward*, *facing*, *held*), along with nuances of attitude or demeanor (*confident*, *relaxed*), and references to upper body parts (*shoulders*, *head*, *stance*).
3. **Cluster 3: Body Structure** – Centers on body form and build, especially the upper body (*body*, *upper*, *torso*), and physical constitution (*slender*, *build*, *average*).
4. **Cluster 4: Accessories & Adornments** – Includes items that complement the body or outfit, such as *jewelry*, *watches*, and *glasses*. Also includes general terms related to adornment or framing (*accessories*, *adornments*, *frame*, *within*).
5. **Cluster 5: Clothing & Garment Details** – Describes clothing items (*shirt*, *jacket*), attributes (*sleeveless*, *thin*, *around*), materials (*fabric*, *material*, *smooth*), colors (*dark*, *white*, *black*), and design elements (*collar*, *neckline*, *design*). Also includes general clothing references (*wearing*, *garment*, *attire*).

6. **Cluster 6: Environment** – Covers elements of the scene or setting, such as *background*, *lighting*, *context*, and *setting*. Includes references to photography or composition (*blurred*, *focus*), possible situations or locations (*indoor*, *formal*), and presence of text or branding (*text*, *details*).
7. **Cluster 7: Facial Features & Expression** – Focuses on the face and its components (*face*, *eyes*, *nose*, *lips*, *eyebrows*, *mouth*), structural attributes (*full*, *arched*), and overall expression or gaze (*expression*, *neutral*).

The themes provided in the prompt seem to be followed by Qwen2.5 in its descriptions. The model also introduces hair as a distinct theme, separate from other related aspects such as facial features. Some cluster boundaries reveal interesting patterns in how the model organizes concepts. For example, the clear separation between hair (Cluster 1) and facial features (Cluster 7) indicates a strong conceptual division between these aspects of appearance. Most concepts make sense within the categorical organization that the cluster algorithm revealed, but it is worth noting that some of these cluster boundaries may be debatable. For example, *adult* and *male* appear in Cluster 1 which seems out of place. Moreover, Cluster 5 contains both *chest* and *earrings*, *chest* being more understandable as it may be strongly related to clothing, but *earrings* seems for fit for cluster 4 . Despite these conceptual relationships, the co-occurrence score gives a very good indication that the model has developed distinct descriptive frameworks for these related, but separate, aspects of appearance.

Gender-related terms appear infrequently across descriptions. This suggests that Qwen2.5 may not organize descriptions around gender categories. When specifically asked if it can determine the gender of the person in the photograph, the model often provides a cautious response, such as: **"Based on the visible features in the image, the person appears to be male. This assessment is based on the short hairstyle, facial structure, and clothing style typically associated with males. However, it's important to note that these are general observations and may not definitively determine gender"**. This careful phrasing may, perhaps, stem from the training process of the language model, reflecting an awareness of the complexities surrounding gender identity.

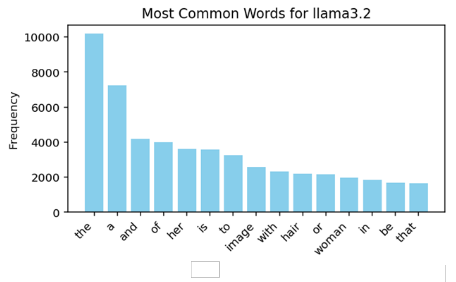
6 Llama3.2-vision:11b Analysis Results

6.1 Word Frequency Analysis

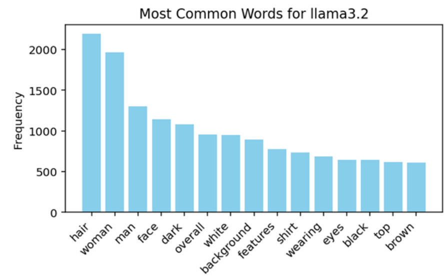
After stop-word removal, the most frequent descriptive terms in Llama3.2's outputs included *hair*, *woman*, *man*, *face*, and *dark* (Figure 3).

6.2 Clustering Analysis

Using the sentence-based co-occurrence method and Louvain community detection algorithm with a resolution parameter of 1.0 on the 100 most frequent



(a) Word frequency distribution before stop word filtering



(b) Word frequency distribution after stop word filtering

Fig. 3: Word frequency distributions before and after stop word filtering Llama3.2's descriptions

words, we identified six distinct clusters with a modularity score of 0.340, illustrated in figure 2. The words in each cluster, ordered by decreasing frequency of occurrence, are the following:

- **Cluster 1:** hair, dark, brown, long, skin, short, back, blonde, straight, depicts, styled, falls, tone, cut, waves.
- **Cluster 2:** woman, overall, presents, attire, appearance, photo, sense, portrait, view, photograph, adds, adding, touch, event, setting, formal, professional, style, taken, elegance, relaxed, atmosphere, clothing, posture.
- **Cluster 3:** man, face, background, features, facial, head, shoulders, subject, expression, left, side, blurred, focus, right, neutral, neck, behind, frame, partially, shoulder, upper, framing, showcasing.
- **Cluster 4:** white, shirt, wearing, black, top, jacket, red, wears, wall, front, collar, backdrop, gray, earrings, purpose, suit, collared.
- **Cluster 5:** eyes, color, blue, camera, smile, lips, subtle, light, looking, eyebrows, directly, small, thin, nose, appear, teeth, smiling, mouth.
- **Cluster 6:** made, smooth, material.

6.3 Findings and Implications

The clustering results from Llama3.2 indicate an organizational framework for person descriptions that appears to differ from that observed in Qwen2.5. These clusters can be interpreted as follows:

1. **Cluster 1: Hair** – Focuses on hair-related traits, including color (*dark*, *blonde*), length (*short*, *long*), texture (*straight*, *waves*), and arrangement (*styled*, *falls*). Also includes references to skin (*skin*, *tone*).
2. **Cluster 2: Presentation & Scene** – Describes the subject's overall presentation and the tone of the image, combining aspects of attire (*clothing*, *attire*), compositional framing (*photo*, *view*), and affective interpretation (*elegance*, *relaxed*, *professional*). Contextual cues (*setting*, *atmosphere*) suggest

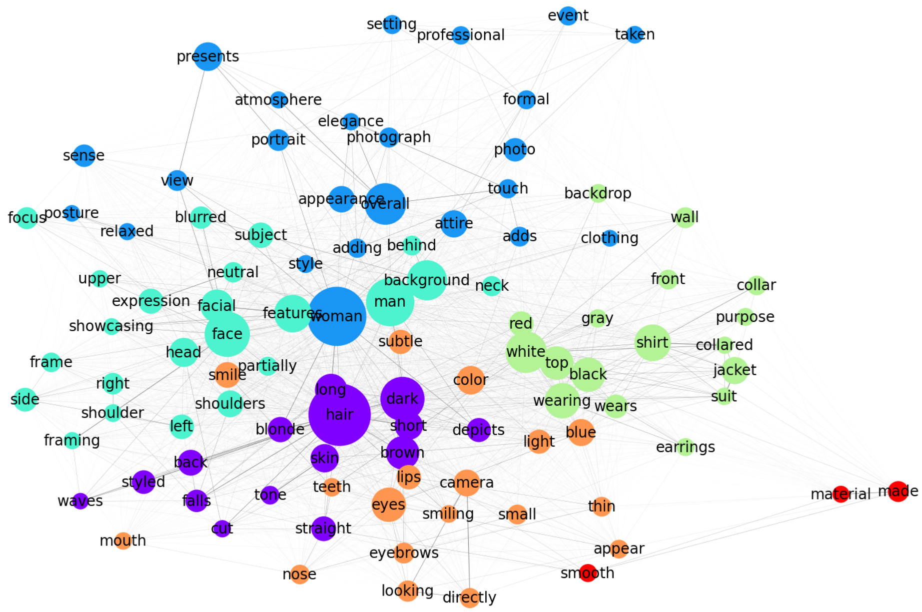


Fig. 4: Graphical representation of lama3.2's word communities, visualized by the Fruchterman-Reingold algorithm[28]

the scene in which the subject appears. Also includes *woman* as a demographic, but the focus is visual composition.

3. **Cluster 3: Framing & Orientation** – Describes how the subject is positioned and framed. Includes body references (*face*, *head*, *shoulders*) and compositional terms (*background*, *framing*). Spatial direction are prominent (*left*, *side*, *behind*), along with affective stance (*expression*, *neutral*). Also includes *man* as a demographic.
4. **Cluster 4: Clothing** – Covers garments (*shirt*, *jacket*), color references (*white*, *black*, *red*), and clothing details (*collar*, *collared*, *earrings*).
5. **Cluster 5: Facial Features & Expression** – Focuses on specific facial elements and emotional expression (*eyes*, *nose*, *smile*). Visual detail terms (*small*, *thin*, *light*) and gaze direction (*looking*, *directly*) are also present.
6. **Cluster 6: Residual community** – A minimal cluster containing a few disconnected terms, likely reflecting residual or spillover content from more coherent clusters.

Like Qwen2.5, the themes outlined in the prompt are largely followed by Llama3.2, and it also introduces hair as a distinct theme. However, unlike Qwen2.5, it largely ignores accessories and instead emphasizes gender, with *woman* and *man* being two of the three most frequently used words.

There seems to more conceptual overlap between clusters, when compared to the Qwen2.5's clusters. Cluster 2 and 3 include *woman* and *max* as their most frequent words, respectivelly, altought they seem to add little to the whole subject of the cluster. Background related words seem spread through different cluster, such as *setting* and *view* being in cluster 2, *background* and *frame* being in cluster 3 and *wall* being in cluster 4. Cluster 3 seems to include themes related to face, body, expression and background, which all seem to deserve their own cluster. Altough the conceptual overlap seems more proeminent than in Qwen2.5's clusters, the co-occurrence score still gives a good indication that the model has distinct descriptive frameworks for these related, but separate, aspects of appearance.

7 Conclusion

Both models largely adhered to the categories suggested in our prompt (facial features, body, posture, clothing, accessories, and other observable details), but with small variations. Both models independently established hair characteristics as a prominent descriptive category separate from facial features. Qwen2.5 demonstrated a more objective, attribute-centered approach with clear distinctions between conceptual categories. In contrast, Llama3.2 exhibited a more gender-centric organization, with *woman* and *man* among its most frequent descriptive terms. Although not explicitly requested, when asked to describe a person, these models failed to introduce themes like tattoos, scars, or birthmarks. The models tended to focus on what was present in the image, largely ignoring what was absent, which can sometimes be just as helpful in providing a more complete description.

The clustering analysis revealed that Qwen2.5 maintained clearer boundaries between conceptual categories, with a modularity score of 0.480 compared to Llama3.2's 0.340. Moreover, the methodological differences required for effective clustering (sentence-based co-occurrence for Qwen2.5 versus sliding window for Llama3.2) may reflect fundamental differences in how these models structure their descriptions.

These findings have important implications for understanding representation in multimodal AI systems. The distinct organizational frameworks identified may reflect underlying biases in training data or architectural differences that influence how these systems perceive and prioritize human attributes. The stronger gender emphasis in Llama3.2's descriptions suggests potential gender-based categorization that could reinforce stereotypes if not carefully addressed.

Future work may expand this analysis to include a wider range of multimodal models with varying architectures and training methodologies to determine whether the patterns observed here are widespread. Longitudinal studies tracking changes in these patterns across model versions would also provide valuable insights into how representation evolves as these technologies advance.

By understanding how LLMs conceptualize and describe people, we can develop more effective interventions to mitigate bias and ensure that these increasingly influential systems represent human diversity fairly and accurately. This work contributes to the growing body of research on responsible AI development and deployment, particularly as these technologies become more deeply integrated into applications that make consequential decisions about human beings.

References

1. Brown, T., 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 preprint arXiv:2005.14165 (2020)
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I.: Attention is All You Need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
3. 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: International Conference on Machine Learning (2021)
4. Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., Kalai, A. T.: Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In: Advances in Neural Information Processing Systems, vol. 29 (2016)
5. Qwen.: Qwen2.5-VL-72B-Instruct. Hugging Face. Accessed February 25, 2025. https://huggingface.co/Qwen/Qwen2.5-VL-72B-Instruct
6. Meta AI.: Llama 3.2. Hugging Face. Accessed February 25, 2025. https://huggingface.co/meta-llama/Llama-3.2-1B
7. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep Learning Face Attributes in the Wild. In: IEEE International Conference on Computer Vision (ICCV), pp. 3730–3738 (2015). https://doi.org/10.1109/ICCV.2015.425
8. Bender, E. M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623 (2021). https://doi.org/10.1145/3442188.3445922
9. Tyler, R., Towler, A., Kemp, R.I., White, D.: Let's talk about faces: Identifying faces from verbal descriptions. British Journal of Psychology, vol. 114, no. 1, pp. 262–281 (2022). https://doi.org/10.1111/bjop.12610
10. Diego-Mas, J.A., Fuentes-Hurtado, F., Naranjo, V., Alcañiz, M.: The influence of each facial feature on how we perceive and interpret human faces. I-Perception, vol. 11, no. 5 (2020). https://doi.org/10.1177/2041669520961123
11. Stolier, R.M., Hehman, E., Keller, M.D., Walker, M., Freeman, J.B.: The conceptual structure of face impressions. Proceedings of the National Academy of Sciences, vol. 115, no. 37, pp. 9210–9215 (2018). https://doi.org/10.1073/pnas.1807222115
12. OpenAI: GPT-4 Technical Report. arXiv, Cornell University (2024). https://doi.org/10.48550/arxiv.2303.08774
13. MetaAI: LLaMA: Open and Efficient Foundation Language Models. arXiv, Cornell University (2023). https://doi.org/10.48550/arxiv.2302.13971
14. Google Research: PaLM: Scaling Language Modeling with Pathways. arXiv, Cornell University (2022). https://doi.org/10.48550/arxiv.2204.02311
15. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. arXiv, Cornell University (2021). https://doi.org/10.48550/arxiv.2102.12092

16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. arXiv, Cornell University (2021). https://doi.org/10.48550/arxiv.2103.00020
17. Sharma, S.: Generating accurate human face sketches from text descriptions. International Journal of Advanced Science Computing and Engineering, vol. 6, no. 1, pp. 20–26 (2024). https://doi.org/10.62527/ijasce.6.1.195
18. Sádaba-Campo, N., Gómez-Moreno, H.: Exploration of generative neural networks for police facial sketches. Big Data and Cognitive Computing, vol. 9, no. 2, p. 42 (2025). https://doi.org/10.3390/bdcc9020042
19. Xiang, C., Xiang, C.: Developers created AI to generate police sketches. Experts are horrified. VICE, February 7, 2023. https://www.vice.com/en/article/ai-police-sketches/
20. Kazemi, H., Iranmanesh, M., Dabouei, A., Soleymani, S., Nasrabadi, N.M.: Facial attributes guided deep sketch-to-photo synthesis. In: IEEE Winter Conference on Applications of Computer Vision Workshops (2018). https://doi.org/10.1109/wacvw.2018.00006
21. European Parliament: EU AI Act: First Regulation on Artificial Intelligence. European Parliament (2025, February 19).
22. Zhao, D., Wang, A., Russakovsky, O.: Understanding and evaluating racial biases in image captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 14830–14840. IEEE (2021). https://doi.org/10.1109/ICCV48922.2021.01456
23. Hamidieh, K., et al.: Identifying implicit social biases in vision-language models. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 547–561. ACM (2024). https://doi.org/10.1145/1234567.8901234
24. Lee, N., et al.: Survey of social bias in vision-language models. arXiv preprint arXiv:2309.14381 (2023). https://arxiv.org/abs/2309.14381
25. Yang, C.: Who's Afraid of George Kingsley Zipf? Or: Do Children and Chimps Have Language? *Significance* **10**(6), 29–34 (2013). https://doi.org/10.1111/j.1740-9713.2013.00708.x
26. Omar, A.: Identifying themes in fiction: A centroid-based lexical clustering approach. *Dil ve Dilbilimi Çalışmaları Dergisi* **2021**, 580–594. https://doi.org/10.17263/jlls.903518
27. Blondel, V.D., et al.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008, 10 (2008). https://doi.org/10.1088/1742-5468/2008/10/P10008
28. Fruchterman, T.M.J., Reingold, E.M.: Graph drawing by force-directed placement. *Software - Practice and Experience* **21**(11), 1129–1164 (1991). https://doi.org/10.1002/SPE.4380211102
29. Traag, V.A., Waltman, L., van Eck, N.J.: From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* **9**(1), 1–12 (2019). https://doi.org/10.1038/s41598-019-41695-z