Extraction and Analysis of Multimodal Concepts in Vision Language Models through Sparse Autoencoders

Sergio Lanza[0009−0004−8777−7070] and Jae Hee Lee[0000−0001−9840−780X]
Stefan Wermter[0000−0003−1343−4775]

Knowledge Technology, Department of Informatics,
University of Hamburg, Hamburg, Germany
sergio.lanza@uni-hamburg.de, jae.hee.lee@uni-hamburg.de,
stefan.wermter@uni-hamburg.de
https://www.inf.uni-hamburg.de/en/inst/ab/wtm.html

Abstract. Vision Language Models (VLMs) have demonstrated impressive performance in tasks requiring joint understanding of images and text, such as image captioning and Visual Question Answering (VQA), but our understanding of their internal processes remains limited. Recently, Sparse Autoencoders (SAEs) have emerged as a promising tool to support the interpretation of concepts encoded in VLMs. However, most SAE-based approaches focus only on textual or visual concepts separately, ignoring multimodal concepts. This limitation hinders a comprehensive understanding of VLMs, since concepts that integrate both modalities can be misclassified. Moreover, previous visual approaches often produce low-quality visual concept descriptions that are vague or incomplete, limiting their usefulness for understanding model reasoning. We propose a framework based on SAEs to extract and analyze visual, textual, and multimodal concepts from VLMs. For each neuron, we propose a candidate human-interpretable concept and compute the alignment between the concept and the dataset samples using cosine similarity scores. Experiments on a VQA dataset (LLaVA-NeXT) demonstrate that our framework improves visual concept quality by up to 45% compared to existing SAE-based methods, while maintaining high textual concept quality and enabling systematic identification of multimodal concepts. This work contributes new insights into the conceptual space of VLMs, providing a structured approach to distinguish between visual, textual, and multimodal concepts. The code is available at https://github.com/PHDLanza/Multidata_SAE.

1 Introduction

Vision Language Models (VLMs) have achieved remarkable results in several tasks from image captioning to Visual Question Answering (VQA), yet understanding their internal mechanisms remains a critical challenge [10,15]. Mechanistic Interpretability (MI) addresses this challenge by studying the internal computation responsible for a model's output [3]. One prominent direction within MI is Concept Extraction, which aims to identify human-understandable concepts encoded

in the model activations to improve model transparency and interpretability [11]. Sparse Autoencoders (SAEs), in particular, have received great attention within Concept Extraction methods for extracting interpretable features encoded in pre-trained models, especially Large Language Models (LLMs) [4,5,6]. SAEs project the high-dimensional embeddings of the LLM into a sparse representation composed of human-understandable features, or concepts, making this approach effective and scalable for concept extraction [3].

However, most SAE-based works focus exclusively on textual concepts, with limited explorations of visual concepts and, to the best of our knowledge, no systematic study of multimodal concepts, i.e., concepts represented jointly in the visual and textual inputs. This gap limits our understanding of VLMs. Recent studies suggest how VLMs encode multimodal information, but rely disproportionately on one modality depending on the task [19]. Understanding which information is encoded by which modality sheds light on the reasoning ability of VLMs and how they align and process inputs across modalities. For example, a neuron may encode a concept that is jointly grounded in the image (an object or region) and in the question-answer text (a property, relation, or action).

In this paper, we present a novel concept extraction framework based on SAEs that integrates visual, textual, and multimodal analysis. For each neuron in the SAE, we select the most representative images and textual samples that most strongly trigger it, then we infer candidate human-interpretable concepts from each modality, and finally, we evaluate which concept best represents the neuron's activation (see Figure 1). The extraction is performed by an external VLM, LLaVA 72B, following prior works on SAE [4,5,7]. The evaluation, instead, is computed using CLIP [18] and ALIGN [9] similarities, which provides a shared space for text-image alignment [17,18].

This work contributes new insights into the concepts encoded inside VLMs, providing a structured approach to distinguish between visual, textual, and multimodal concepts. Our novel contribution is divided into two main components:

- **Integrated multimodal concept extraction:** We extend SAE analysis to extract visual, textual, and multimodal concepts in a unified framework, maintaining high concept quality across the three modalities.
- **Improved visual concept quality:** We introduce a new visual concept extraction technique that yields more precise and semantically grounded visual concepts. Our results on a VQA dataset, LLaVA-NeXT, show an improvement of up to 45% in cosine similarity alignment between concept descriptions and their images compared to the existing SAE method.

2 Background

Sparse Autoencoders (SAEs) [5] have been used to discover concepts encoded in the pretrained models. Specifically, it is an autoencoder that consists of a single high-dimensional hidden-layer (from 2× to over 100× the activation dimension)

trained to reconstruct the activation from the model of interest. During the training phase, SAEs incentivize sparsity in the hidden layer, i.e., they minimize the number of activated neurons inside the layer. Models that enhance sparsity help encode neurons representing a single, distinct concept called monosemantic neurons, which faithfully represent the backbone computation [6].

As defined by Pach et al. [15], an SAE consists of two linear layers:

$$W_{\text{enc}} \in \mathbb{R}^{d \times \omega}, \quad W_{\text{dec}} \in \mathbb{R}^{\omega \times d},$$

as encoder and decoder, respectively. The ω term represents the dimension of the SAE hidden layer defined as $\omega := d \cdot \varepsilon$, where ε is the expansion factor and d the embedding dimension of the hosted model. Given an input $x \in \mathbb{R}^d$, the SAE computes the latent activation $z \in \mathbb{R}^\omega$ and reconstruction $\hat{x} \in \mathbb{R}^d$ as:

$$\begin{aligned} z &= \text{TopK}\big(\text{ReLU}\big(W_{\text{enc}}^\top (x - b)\big)\big), \\ \hat{x} &= W_{\text{dec}}^\top z + b, \end{aligned}$$

where TopK is the activation function that retains the highest K (a hyperparameter) values in the latent neurons while zeroing out all the others, enforcing sparsity in the latent representation, and b is the bias term. The SAE is trained to minimize the reconstruction loss:

$$\mathcal{L}(x) = \|x - \hat{x}\|_2$$

encouraging faithful reconstruction of the original embedding based on the SAE neurons. Since the hidden layer of an SAE can comprise more than 100k neurons, it is infeasible to manually associate each neuron with a concept based on the triggered inputs. The most commonly used approach involves collecting dataset samples (texts or images) that activate the same neuron and then querying an LLM, denoted as explainer, to infer a common concept [4,5,7].

3 Related Work

3.1 Sparse AutoEncoders for VLMs

The success of SAEs in the Language context has recently affected the concept extraction research for VLMs. Particularly, the approach of Zhang et al. [21] automatically detects and extracts concepts from individual SAE neurons injected into a VLM. For each neuron in the latent space of the SAE, the top 5 images with the highest average activations are forwarded to the explainer, LLaVA 72B, [13] that performs a zero-shot concept extraction task. To ensure that the explainer focuses solely on the most interesting regions, they apply a masking process on the image patches: only the patches that activate the selected neuron are shown to LLaVA-72B, while all others are masked in black. A related line of research studied the layers in the CLIP embedding [12,15] using a similar masking technique and exploiting CLIP's semantic labelling capabilities to assess concepts within subpatches of the image. However, all these works do not consider textual concepts

or analyze them superficially, limiting the insight that users can obtain, although all of them suggest the possible presence of these concepts in the VLM [12,15,21]. Moreover, the masking process removes too much contextual information, limiting the interpretability and reducing the accuracy of concepts extracted. Our work addresses these limitations by jointly extracting and evaluating visual, textual, and multimodal concepts within a single SAE-based framework and improving the extraction process with additional textual information (see Section 4.1).

4 Methodology

4.1 Multimodal Concept Extraction

Inspired by previous work [12,21], we design our novel concept hypothesis extractor leveraging an external VLM as an explainer. All experiments are conducted on a VQA dataset; VQA is a task where a model answers a question about an image and is evaluated against a ground-truth answer. First, we feed a set of VQA samples into a VLM augmented with a pretrained SAE to record which sample strongly activates the studied SAE neurons. For each VQA sample j, which comprises an image and a textual question–answer pair, we record two separate SAE activation tensors corresponding to visual and textual tokens defined as:

$$V^{(j)} = [v_1^{(j)}, \ldots, v_P^{(j)}] \in \mathbb{R}^{P \times \omega} , \qquad T^{(j)} = [t_1^{(j)}, \ldots, t_{N_j}^{(j)}] \in \mathbb{R}^{N_j \times \omega}$$

where $v_i^{(j)}$ and $t_i^{(j)}$ are the activation vectors for the i-th visual patch or textual token respectively, P represents the number of visual patches per image (24×24), N_j corresponds to the number of textual tokens encoded in the j-th sample and ω denotes the hidden dimension of SAE. From these matrices, we compute the average activations normalized for patches, $\bar{V}^{(j)} \in \mathbb{R}^{\omega}$, or tokens, $\bar{T}^{(j)} \in \mathbb{R}^{\omega}$, defined as:

$$\bar{V}^{(j)} = \frac{1}{P} \sum_{i=1}^{P} v_i^{(j)}, \qquad \bar{T}^{(j)} = \frac{1}{N_j} \sum_{i=1}^{N_j} t_i^{(j)}$$

For each SAE neuron k, we then select the top-5 samples that maximize $\bar{V}_k^{(j)}$ (visual) and $\bar{T}_k^{(j)}$ (textual), and query the explainer to extract the hypothesis. As illustrated in Figure 1, in the visual extraction process, we apply a black mask to the patches that do not activate the targeted neuron, following prior studies [15,21], while including the corresponding VQA question and answer alongside the image input. We introduce this element following evidence that additional related text improves visual understanding accuracy in VLMs [14,20]. For textual extraction, we apply a similar procedure: we highlight the specific tokens that trigger the neuron by enclosing them in brackets and pass these sentences to the external VLM, including the corresponding VQA images. This approach mirrors the visual process and provides richer multimodal grounding in the textual extraction.

4.2 Metrics

After the extraction of the concept hypotheses (see Section 4.1), each SAE neuron yields up to two hypotheses: visual, textual, or no hypothesis if the available data are insufficient to produce a shared concept.

To quantitatively evaluate the hypotheses, we require a precise and modality-independent measure. We therefore adopt CLIP ViT-B/32 [18] and ALIGN [9] as our evaluators, two vision-language models trained to align images and text in a shared embedding space, which have been used in prior interpretability works [21,15]. We utilize CLIP and ALIGN in two evaluation settings.

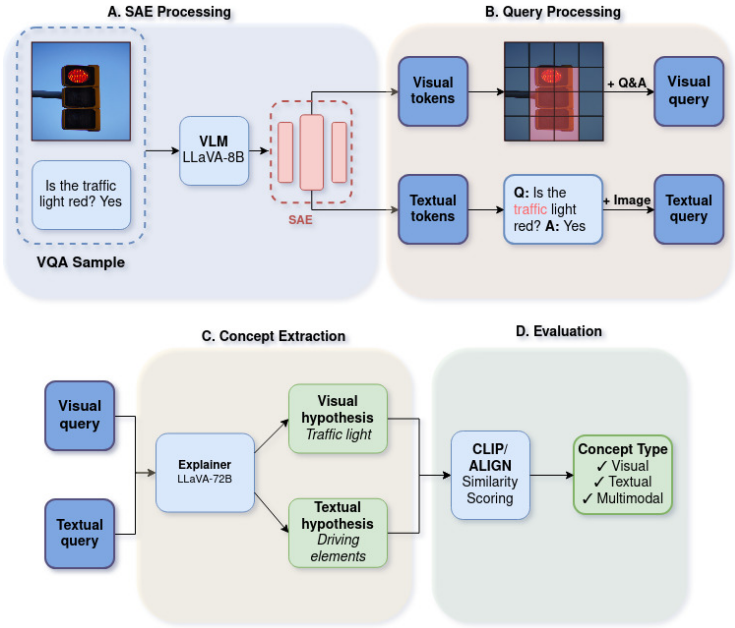

Fig. 1: Overview of our multimodal concept extraction framework. **(A)** VQA sample is processed through a VLM with an integrated SAE that projects activations into a sparse hidden layer, producing visual and textual tokens. **(B)** For a target SAE neuron, we select the visual and textual tokens that strongly activate it and construct modality-specific inputs: non-activating visual tokens (or patches) are black masked, while activating textual tokens are highlighted. From these inputs, we construct two queries, each augmented with complementary information (the question–answer pair for visual extraction and the unmodified image for textual extraction). **(C)** An external VLM (the explainer) extracts concept hypotheses from the queries. **(D)** CLIP/ALIGN evaluates hypothesis data alignment to classify each neuron as visual, textual, or multimodal.

Hypothesis–data alignment In the first setting, for each hypothesis and each evaluator, we input to the model the same data used in the extraction, specifically blurred image patches for visual hypotheses and token sequences for textual hypotheses, and compute the similarity score for all the hypotheses. In preliminary experiments, we observed that black masking degraded the CLIP and ALIGN reliability; therefore, we blurred the inactive patches, leading to more stable and interpretable outcomes. We obtain a similarity matrix of scores for each modality comparing the sets of data used in the extraction process and all the hypotheses, and we rank the highest score for each row. This procedure assesses the alignment between the hypothesis and the data from which it was generated. To account for semantic similarity among hypotheses, we consider a hypothesis validated if its score on the data that generated it ranks within the top n predictions. In our experiments, we use $n \in \{1, 2, \ldots, 5\}$.

Multimodal neuron assignment For neurons with both a visual hypothesis h_{vis} and a textual hypothesis h_{text}, we assign a modality label based on semantic consistency between the two hypotheses. We encode h_{vis} and h_{text} using the text encoders of CLIP and ALIGN and compute cosine similarity. We label a neuron as *multimodal* if both similarities exceed a threshold τ ($\tau = 0.8$ in our experiments); otherwise, we assign it to the modality whose hypothesis yields the higher hypothesis–data alignment score. In Section 5.4, we report a sensitivity analysis over other τ values.

Category-based evaluation For the second setting, we follow the evaluation procedure of Zhang et al. [21] to ensure a direct comparison with our baseline. Each visual hypothesis is mapped to a specific visual category, and then we compute the CLIP and ALIGN scores between the samples and the assigned category label. An external VLM performs the category mapping, and the six categories follow the taxonomy introduced in [2]. While the first setting offers a more fine-grained evaluation and explicitly accounts for textual and multimodal concepts, we also adopt the second to enable a fair comparison with our baseline (see Section 5), which focuses solely on visual concepts.

Additional textual evaluation Finally, we assess the quality of textual extraction to verify that textual hypotheses are not biased by the inclusion of visual information during the extraction procedure. Following prior works described in Section 2, we query an external LLM explainer with randomly sampled textual sentences from LLaVA-NeXT together with the analyzed hypothesis. The explainer predicts whether the hypothesis is effectively represented in the sentences, and then we compute the accuracy and the precision against the ground-truth annotations. We utilize the evaluation protocol of Paulo et al. [17], specifically, we adopt the detection score, which evaluates hypothesis presence at the sentence level, and the fuzzing score, focused on hypothesis presence at the token level.

5 Experiments

In this section, we present the details of the quality and modality distribution of extracted concepts. We compared our results to Zhang et al. [21], referring to it as our baseline. To be consistent with the baseline, our analysis focuses only on the first 5000 neurons of the SAE.

5.1 Models

For consistency with prior work [21], we adopt the same VLMs and SAE. The SAE is trained on frozen embeddings extracted from the LLaVA-NeXT-8B-HF model [13] with Top-k activation ($k = 256$) and a latent space $\omega = 2^{17}$ neurons. Based on prior work, we select the SAE trained on the 24^{th} transformer layer, which empirically yields the highest visual interpretability. For concept hypothesis extraction as an explainer, we adopt LLaVA-72B and select the five best samples to remain consistent with our baseline. For the textual analysis only, we utilize Llama-3.1-70B-Instruct [8], an LLM already adopted in another textual SAE analysis work [17] to avoid undesired visual bias that the VLM could bring. All experiments are run on two NVIDIA H100 GPUs: a complete run of the framework, from activation extraction to concept evaluation, requires approximately 30 hours of computation.

Table 1: Comparison of correct visual concept extraction rates between our framework (**Ours**) and the **baseline** [21] using CLIP and ALIGN scores. Our method outperforms the baseline at all thresholds by an average margin of 41% (CLIP) and 45% (ALIGN).

Thresh.	**Base (CLIP)**	**Ours (CLIP)**	**Base (ALIGN)**	**Ours (ALIGN)**
	Avg (↑)	Avg (↑)	Avg (↑)	Avg (↑)
1	0.15	**0.45**	0.13	**0.47**
2	0.22	**0.61**	0.20	**0.65**
3	0.26	**0.69**	0.24	**0.69**
4	0.29	**0.75**	0.26	**0.74**
5	0.31	**0.79**	0.27	**0.80**
Avg	0.246	**0.658**	0.220	**0.670**

5.2 Datasets

To remain consistent with our visual baseline, our main experiment uses the LLaVA-NeXT dataset [13], which follows the VQA format: each sample comprises an image, a corresponding question, and an answer. We also replicate our experiments on a subset of the COCO dataset [16] to evaluate the generalizability of

our approach with a different set of data (see Appendix A). Due to computational constraints and to remain consistent with our baseline, we use a subset of it, specifically the first 15% of LLaVA-NeXT [21]. All prompts used are reported in the Appendix C.

5.3 Multimodal Extraction

In Table 3, we show the number of concept hypotheses extracted by our framework and the baseline (see Appendix B for some qualitative examples). Although our framework extracts fewer visual concepts compared to the baseline, we argue that the concepts identified by our approach are more semantically representative of the images utilized to extract them. To validate this claim, we compute the direct concept alignment scores following the procedures described in Section 4.2. As illustrated in Table 1, our approach consistently outperforms the baseline with higher accuracy in all thresholds and in both metrics, indicating a stronger alignment between images and the identified visual and multimodal concepts. To confirm this, Figure 2 shows that our approach identifies fewer concepts with low CLIP scores compared to the baseline. Instead, in Table 2, our framework

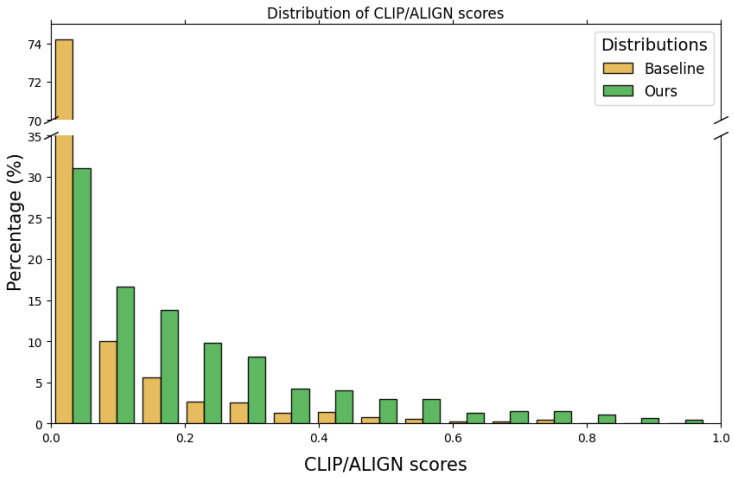

Fig. 2: Distribution of the CLIP and ALIGN scores for the proposed model (Ours) compared to the baseline. For each sample, we compute the average between the CLIP and ALIGN scores. The histogram illustrates the percentage of samples across similarity score intervals, highlighting a shift toward higher average similarity values in the proposed method relative to the baseline distribution.

achieves comparable or slightly superior results in the category-concept alignment score. Finally, in Table 4 we assessed the quality of textual concepts using the

Interpretability score [17]. We compare our approach with other similar SAEs using only the textual approach, obtaining similar results.

Table 2: Average accuracy of correct visual concept categorization through CLIP and ALIGN similarities with [2] labels.

Label	**Baseline (CLIP)**		**Ours (CLIP)**		**Baseline (ALIGN)**		**Ours (ALIGN)**	
	Avg (↑)	Var (↓)	Avg (↑)	Var (↓)	Avg (↑)	Var (↓)	Avg (↑)	Var (↓)
scene	18.84	9.4×10^{-3}	**26.58**	1.3×100^{-2}	16.84	1.5×10^{-6}	**16.94**	1.4×10^{-6}
object	**40.65**	2.3×10^{-2}	39.02	2.2×10^{-2}	16.69	1.4×10^{-6}	**16.73**	1.4×10^{-6}
part	**23.24**	8.7×10^{-3}	20.93	5.0×10^{-3}	**16.69**	1.0×10^{-6}	16.66	1.8×10^{-6}
material	**5.36**	1.0×10^{-3}	4.98	6.0×10^{-4}	**16.59**	1.0×10^{-6}	16.57	1.8×10^{-6}
texture	4.69	3.0×10^{-3}	**9.61**	8.4×10^{-3}	16.69	1.9×10^{-6}	**16.86**	1.8×10^{-6}
color	**32.27**	2.8×10^{-2}	25.76	2.1×10^{-2}	**16.82**	2.1×10^{-6}	16.69	2.0×10^{-6}
Avg	20.84	1.2×10^{-2}	**21.98**	1.1×10^{-2}	16.72	1.4×10^{-6}	**16.74**	1.6×10^{-6}

Table 3: Number of concept hypotheses from our framework and the baseline [21].

Modality	**Zhang**	**Ours**
Textual	–	2291
Visual	1391	470
Multimodal	–	132

Table 4: Interpretability scores for textual concepts [17]. Average of 5 runs.

Models	**Det.**	**Fuzz.**
Gemma 2 9b	0.76	0.74
Llama 3.1 8b	0.81	0.83
LLaVA (Ours)	0.78	0.80

Table 5: Multimodal concept counts at different thresholds (249 SAE neurons).

Threshold	**Count**
0.70	223
0.80	132
0.90	23

5.4 Analysis for multimodal threshold

In this section, we analyze the effect of different values of the threshold τ. Table 5 reports the number of neurons whose visual and textual hypotheses exceed each threshold for both CLIP and ALIGN. As expected, increasing τ results in a stricter criterion and, as a consequence, a smaller number of multimodal concepts. Conversely, decreasing τ increases the number of multimodal concepts. Based on this analysis, we select an intermediate threshold of $\tau = 0.8$ which balances between coverage and semantic consistency.

5.5 Qualitative Intervention

In this experiment, we want to evaluate the impact of specific concepts on text generation qualitatively [15]. We tested the LLaVA-NeXT-8B-HF model, already

integrated with our SAE, with the instruction "**Tell me a story about a person**". The prompt is provided in two versions: one with an associated image as additional input and one without, to observe how visual information affects the intervention. During the new generation, we force the target neuron's activation to the fixed value $\alpha = 60$, nearly 20 times the average activation of an SAE neuron, following past results [15]. We select three concepts, one from each modality, to analyze which modality has the biggest impact on text generation. As illustrated in Figure 3, the output modifications are more pronounced generally when the input includes an image, suggesting that visual features amplify the effect of concept-level intervention, and meanwhile multimodal concepts seem to affect the output production more than the other two categories.

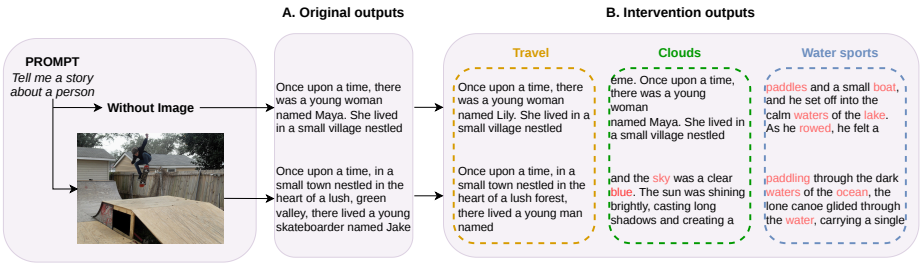

Fig. 3: Qualitative example for intervention experiments using the prompt "**Tell me a story about a person**". **(A)** We generate two outputs, with and without an image prompt. **(B)** We show six intervention outputs, where three concepts are injected independently for each prompt. The concepts are selected from the three categories: textual, visual, and multimodal. Words referring to the injected concept are highlighted in red.

6 Discussion

Our work primarily focuses on a single SAE and a single VLM as an explainer. Future work will extend the generalizability of the proposed framework across multiple architectures. Our concepts are generated from two independent prompts, designed to produce concise and interpretable descriptions rather than longer, free-form explanations. While this choice improves consistency and evaluation reliability, particularly because CLIP and ALIGN are more effective on short-medium length texts, it may restrict the diversity and semantic richness of the extracted concepts. Finally, our findings reveal a clear correlation between SAE's concept alteration with the text output, but they do not establish a causal relationship.

7 Conclusion

In this paper, we demonstrate that an SAE encodes textual and visual multimodal concepts condensed into a shared latent space. Based on this property, we design a novel framework that performs dual-modality concept extraction, enriching both visual and textual hypotheses with multimodal information. Our results exhibit a significant improvement in the quality of interpretable SAE neurons connected to visual concepts, achieving an average improvement of 41% and 45% for CLIP and ALIGN scores, maintaining good quality textual production, and providing a more comprehensive understanding of the internal reasoning of VLMs. Finally, our intervention studies confirm the influence of multimodal and visual concepts on text generation. Overall, this work offers a new understanding of the concept space representation of VLMs by leveraging a new SAE-based approach. We hope this work encourages future research on scalable SAE integration, multimodal interpretability, and concept-guided control in multimodal reasoning systems.

8 Acknowledgments

The authors gratefully acknowledge funding from Horizon Europe under the MSCA grant agreements No 101072488 (TRAIL),No 101226624 (GREET), and No 101168792 (SWEET), as well as from the German Research Foundation (DFG), project number 551629603.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual Question Answering. In: International Conference on Computer Vision (ICCV) (2015)
2. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3319–3327. Honolulu, HI (Jul 2017). https://doi.org/10.1109/CVPR.2017.354, https://ieeexplore.ieee.org/document/8099837
3. Bereska, L., Gavves, S.: Mechanistic interpretability for AI safety - a review. Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=ePUVetPKu6
4. Bricken, Trenton, Templeton, Adly, Batson, Joshua, Chen Brian, Jermyn Adam: Towards monosemanticity: decomposing language models with dictionary learning (Oct 2023), https://transformer-circuits.pub/2023/monosemantic-features
5. Cunningham, H., Ewart, A., Riggs, L., Huben, R., Sharkey, L.: Sparse autoencoders find highly interpretable features in language models. arXiv preprint arXiv:2309.08600 (Oct 2023). https://doi.org/10.48550/arXiv.2309.08600, https://arxiv.org/abs/2309.08600
6. Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, e.a.: Toy models of superposition. arXiv preprint arXiv:2209.10652 (Sep 2022). https://doi.org/10.48550/arXiv.2209.10652, https://arxiv.org/abs/2209.10652

7. Gao, L., la Tour, T.D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., Wu, J.: Scaling and evaluating sparse autoencoders. arXiv preprint arXiv:2406.04093 (Jun 2024). https://doi.org/10.48550/arXiv.2406.04093, https://arxiv.org/abs/2406.04093
8. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., et al.: The llama 3 herd of models (2024), https://arxiv.org/abs/2407.21783
9. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. vol. 139, pp. 4904–4916 (18–24 Jul 2021), https://proceedings.mlr.press/v139/jia21b.html
10. Kaduri, O., Bagon, S., Dekel, T.: What's in the image? a deep-dive into the vision of vision language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14549–14558 (June 2025), https://openaccess.thecvf.com/content/CVPR2025/html/Kaduri_Whats_in_the_Image_A_Deep-Dive_into_the_Vision_of_CVPR_2025_paper.html
11. Lee, J.H., Lanza, S., Wermter, S.: From neural activations to concepts: A survey on explaining concepts in neural networks (2024), https://neurosymbolic-ai-journal.com/system/files/nai-paper-697.pdf
12. Lim, H., Choi, J., Choo, J., Schneider, S.: Sparse autoencoders reveal selective remapping of visual concepts during adaptation. arXiv preprint arXiv:2412.05276 (Mar 2025). https://doi.org/10.48550/arXiv.2412.05276, https://arxiv.org/abs/2412.05276
13. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/
14. Ma, P., Rietdorf, L., Kotovenko, D., Hu, V.T., Ommer, B.: Does vlm classification benefit from llm description semantics? In: Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (2025). https://doi.org/10.1609/aaai.v39i6.32638, https://doi.org/10.1609/aaai.v39i6.32638
15. Pach, M., Karthik, S., Bouniot, Q., Belongie, S., Akata, Z.: Sparse autoencoders learn monosemantic Features in vision-language models. Poster on Neural Information Processing Systems (NeurIPS 2025) (Apr 2025). https://doi.org/10.48550/arXiv.2504.02821, https://neurips.cc/virtual/2025/loc/san-diego/poster/119210
16. Park, D.H., Hendricks, L.A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., Rohrbach, M.: Multimodal explanations: Justifying decisions and pointing to the evidence. CoRR (2018), http://arxiv.org/abs/1802.08129
17. Paulo, G., Mallen, A., Juang, C., Belrose, N.: Automatically interpreting millions of features in large language models. arXiv preprint arXiv:2410.13928 (Dec 2024). https://doi.org/10.48550/arXiv.2410.13928, https://arxiv.org/abs/2410.13928
18. 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. In: Proceedings of the 38th International Conference on Machine Learning. vol. 139, pp. 8748–8763 (18–24 Jul 2021), https://proceedings.mlr.press/v139/radford21a.html
19. Sim, M.Y., Zhang, W.E., Dai, X., Fang, B.: Can VLMs actually see and read? a survey on modality collapse in vision-language models. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Findings of the Association for Computational Linguistics:

ACL 2025. pp. 24452–24470. Association for Computational Linguistics, Vienna, Austria (Jul 2025). https://doi.org/10.18653/v1/2025.findings-acl.1256, https://aclanthology.org/2025.findings-acl.1256/
20. Zang, Y., Yun, T., Tan, H., Bui, T., Sun, C.: Pre-trained vision-language models learn discoverable visual concepts. Transactions on Machine Learning Research (2025), https://openreview.net/forum?id=Vq0wMFBjo2
21. Zhang, K., Shen, Y., Li, B., Liu, Z.: Large multimodal models can interpret features in large multimodal models. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (Nov 2025), https://openaccess.thecvf.com/content/ICCV2025/papers/Zhang_Large_Multi-modal_Models_Can_Interpret_Features_in_Large_Multi-modal_Models_ICCV_2025_paper.pdf

A Analysis on COCO dataset

In this section, we report additional results of our framework applied to a different VQA dataset based on COCO[1], using the SAE injected into LLaVA-NeXT-8B-HF model. Although SAEs are generally not conditioned on the dataset used during inference, with this test, we verify whether our framework preserves this property for the first 1000 neurons.

Table 6: COCO generalization results for the first 1000 neurons.

Dataset	**#Text**	**#Visual**	**#Multimodal**
LLaVA-NeXT (main)	639	85	19
COCO (subset)	537	57	9

We also compare concept hypotheses across datasets using CLIP and ALIGN embeddings to estimate the overlap and stability of extracted concepts.

Table 7: Average similarity scores between the two dataset

Metric	**LLaVA-NeXT**	**COCO**
CLIP	639	85
ALIGN	537	57

B Qualitative Examples

In this section, we present qualitative examples of concepts grouped by modality (Figures 4, 5, and 6):

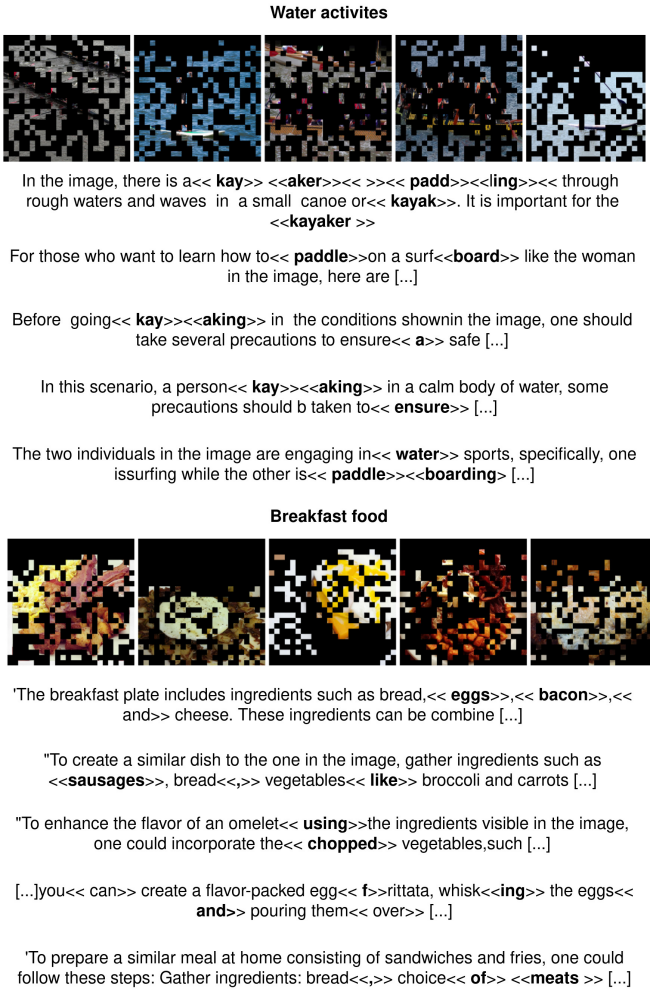

Fig. 4: Examples of multimodal concepts. The textual tokens that trigger the SAE neuron are highlighted with «».

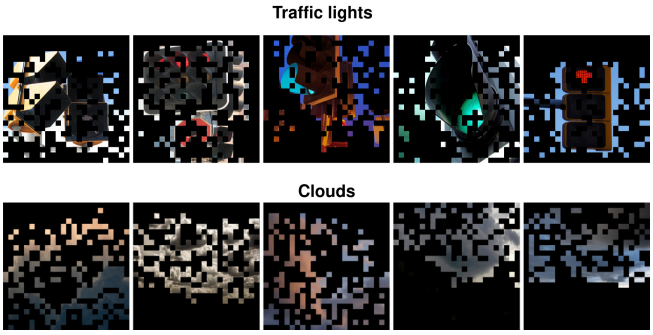

Fig. 5: Examples of visual concepts.

Preparation and Strategy in Sports

During a baseball game, an on-deck batter may analyze their teammate's at-bat to gather useful insights and strategize their approach<< **when**>> it's their<< **turn**>> bat. By observing the opposing pitcher's technique

The next course of action for the player warming up, who is standing near the catcher and the umpire, would be to prepare himself for<< **turn**>> bat. It's likely that he is observin

"While the boy in the baseball uniform is swinging his baseball bat on the field, the other players waiting for<< **turn**>> can take advantage of the opportunity to observe his technique and learn from it.

Black and white photography

'Based on the<< **black**>><< **and**>> white nature of the photo depicting a baseball team, it can be inferred that the era when this photo was taken is likely to be in the early-to-mid 20th century. and>> white[..]

The choice of black<< and>> white for this image of a person surfing might have been made for several reasons. \n\nFirstly, black<< **and**>> white photography [..]

One possible reason for the black<< **and**>> white photograph could be that the image is an older photograph [..]

Fig. 6: Examples of textual concepts.The textual tokens that trigger the SAE neuron are highlighted with «».

C Prompts

Below are the prompts used to query the LLaVA-72B model (Figures 7, 8, 9), inspired by or directly taken from [21]. The first prompt maps each visual hypothesis to one of six predefined labels, and the next two prompts generate visual and textual hypotheses (see Section 4.1).

[GUIDELINES]
You are an AI assistant tasked with assigning a single label based on the given input text. Each input will contain a description of a visual feature, which you must categorize into one of the following classes:
scene - Describes a scene or environment.
object - Describes an object or entity.
part - Describes a part or aspect of an object.
material - Describes a material or substance that constitutes other objects.
texture - Describes the texture of an object.
color - Describes the color of an object.
Please provide only the class label from the list [scene, object, part, material, texture, color] with no additional text.
Only one label should be chosen. Make sure you only choose from the classes listed above and do not output any other classes.
Categorize the following description: {description}
ANSWER:

Fig. 7: Labelling Prompt (LLaVA 72B)

The last two prompts are used to query Llama 3.1 70B, which evaluates textual concepts (Figures 10 and 11) based on [17].

[REQUIREMENTS]
Focus only on the highlighted region in each image. If no region is highlighted or if the highlighted region is minimal (e.g., a few bright spots), ignore the image. Identify common visual patterns, objects, or concepts in the activated regions. For example, note if highlighted areas show consistent structures, such as mesh patterns or similar objects.
[GUIDELINES]
1. Consider Text Context: While maintaining primary focus on the highlighted regions in images, you may marginally consider the associated text (questions and answers) to support or refine your visual observations. However, the final concept should be predominantly based on visual patterns.
2. Concise Description Only: Provide a short, direct description of the common features within the highlighted regions. Avoid any interpretive language—simply state what you see, such as “mesh-like structures” or “actions related to joy or happiness”.
3. Describe Only the Highlighted Regions: Generate captions solely based on the highlighted regions. If no meaningful pattern is visible, or if only a few scattered spots are highlighted, output: "Concept: ‘No visual concept‘"

Fig. 8: Visual Generation Guidelines (LLaVA 72B)

[REQUIREMENTS]
Focus only on the text content provided with each example. If the text is missing, irrelevant, or extremely minimal (e.g., a few unrelated words), ignore that example.
Identify common themes, objects, or concepts mentioned across the text snippets. Pay special attention to any highlighted word in each text—this word should be treated as the most important cue for concept identification.
[GUIDELINES]
1. You will receive a series of text snippets, sometimes accompanied by images. Only use the text, and in particular the word between parentheses, to identify the shared concept. Images should not be considered in your analysis.
These examples are derived from a Visual Question Answering dataset, so each text is in the form of a question or an answer.
2. Concise Description Only: Provide a short, direct description of the common concept emerging from the texts. Avoid speculation or abstract interpretation—simply state what is explicitly or implicitly repeated, especially in relation to the highlighted words (e.g., “vehicles,” “cooking actions,” “types of animals”). Use the image only for reference if absolutely necessary; the main analysis must be text-driven, with words in parentheses as priority.
3. If no clear concept emerges from the texts (e.g., if they are too diverse or vague), write: No textual concept
[OUTPUT EXAMPLES]
Concept: "A tennis match"
Concept: "Descriptions of birds"
Concept: "No textual concept"
Remember, Write always only one Concept for the entire set of inputs

Fig. 9: A4: Textual Generation Guidelines (LLaVA 72B)

"""You are an intelligent and meticulous linguistics researcher.
You will be given a certain latent of text, such as "male pronouns" or "text with negative sentiment".
You will then be given several text examples. Your task is to determine which examples possess the latent.
For each example in turn, return 1 if the sentence is correctly labeled or 0 if the tokens are mislabeled. You must return your response in a valid Python list. Do not return anything else besides a Python list.
Latent explanation: Words related to American football positions, specifically the tight end position.
Test examples:
Example 0:<|endoftext|>Getty ImagesĊĊPatriots tight end Rob Gronkowski had his bossâĢĻ Example 1: names of months used in The Lord of the Rings:ĊĊâĢľâĢ¦the Example 2: Media Day 2015ĊĊLSU defensive end Isaiah Washington (94) speaks to the Example 3: shown, is generally not eligible for ads. For example, videos about recent tragedies, Example 4: line, with the left side âĢĶ namely tackle Byron Bell at tackle and guard Amini
"[1,0,0,0,1]"
Latent explanation: The word "guys" in the phrase "you guys".
Test examples:
Example 0: enact an individual health insurance mandate?âĢŁ., Pelosi's response was to dismiss both Example 1: birth control access<|endoftext|> but I assure you women in Kentucky aren't laughing as they struggle Example 2: du Soleil Fall Protection Program with construction requirements that do not apply to theater settings because Example 3:Ċ<|endoftext|> distasteful. Amidst the slime lurk bits of Schadenfre Example 4: the<|endoftext|>ľI want to remind you all that 10 days ago (director Massimil
"[0,0,0,0,0]"
Latent explanation: "of" before words that start with a capital letter.
Test examples:
Example 0: climate, TomblinâĢĻs Chief of Staff Charlie Lorensen said.Ċ Example 1: no wonderworking relics, no true Body and Blood of Christ, no true Baptism Example 2:ĊĊDeborah Sathe, Head of Talent Development and Production at Film London, Example 3:ĊĊIt has been devised by Director of Public Prosecutions (DPP) Example 4: and fair investigation not even include the Director of Athletics? A· Finally, we believe the
"[1,1,1,1,1]" """

Fig. 10: Detection score prompt (Llama 3.1 70B)

""" You are an intelligent and meticulous linguistics researcher.
You will be given a certain latent of text, such as "male pronouns" or "text with negative sentiment". You will be given a few examples of text that contain this latent. Portions of the sentence which strongly represent this latent are between tokens « and ».
Some examples might be mislabeled. Your task is to determine if every single token within « and » is correctly labeled. Consider that all provided examples could be correct, none of the examples could be correct, or a mix. An example is only correct if every marked token is representative of the latent
For each example in turn, return 1 if the sentence is correctly labeled or 0 if the tokens are mislabeled. You must return your response in a valid Python list. Do not return anything else besides a Python list. Latent explanation: Words related to American football positions, specifically the tight end position.
Test examples:
Example 0:<|endoftext|>Getty ImagesĊĊPatriots« tight end» Rob Gronkowski had his bossâĢĻ Example 1: posted<|endoftext|>You should know this« about» offensive line coaches: they are large, demanding« men» Example 2: Media Day 2015ĊĊLSU« defensive» end Isaiah Washington (94) speaks« to the» Example 3:« running backs»," he said. .. Defensive« end» Carroll Phillips is improving and his injury is Example 4:« line», with the left side âĢĶ namely« tackle» Byron Bell at« tackle» and« guard» Amini
"[1,0,0,1,1]"
Latent explanation: The word "guys" in the phrase "you guys".
Test examples:
Example 0: if you are« comfortable» with it. You« guys» support me in many other ways already and Example 1: birth control access<|endoftext|> but I assure you« women» in Kentucky aren't laughing as they struggle Example 2:âĢĻs gig! I hope you guys« LOVE» her, and« please» be nice, Example 3:American, told« Hannity» that âĢľyou« guys» are playing the race card.âĢŁ. Example 4:« the»<|endoftext|>ľI want to« remind» you all that 10 days ago (director Massimil
"[0,0,0,0,0]"
Latent explanation: "of" before words that start with a capital letter.
Test examples:
Example 0: climate, TomblinâĢĻs Chief« of» Staff Charlie Lorensen said.Ċ Example 1: no wonderworking relics, no true Body and Blood« of» Christ, no true Baptism Example 2:ĊĊDeborah Sathe, Head« of» Talent Development and Production at Film London, Example 3:ĊĊIt has been devised by Director« of» Public Prosecutions (DPP) Example 4: and fair investigation not even include the Director« of» Athletics? A· Finally, we believe the
"[1,1,1,1,1]" """

Fig. 11: Fuzzing score prompt (Llama 3.1 70B)