Language Models Can Explain Visual Features via Steering

Supplementary Material

A. SAE Training Details

For training the SAEs, we used the `dictionary_learning` library [31]. All SAEs were optimized using the Adam optimizer with a learning rate of 3×10^{-4}, $\beta_1 = 0.9$, and $\beta_2 = 0.99$. Training was conducted over a single epoch of the ImageNet training set (1.28M images) with a batch size of 8192. We enforced a sparsity constraint of 25 active features per patch position.

Model activations from `HuggingFace` [52] were cached on-the-fly during training. We maintained a buffer of 500 million activations, from which we randomly sampled. When the buffer was depleted to half capacity, it was refilled with new activations.

Table 3. Explanation evaluation metrics for the later layer SAE of Gemma 3 vision encoder. Except for AUROC, mean scores are reported, and statistical significance is assessed pairwise between methods. A value is underlined if it is significantly higher (with $p < 0.05$) than both other methods in the same column.

Explanation Method	**IoU Score**		**AUROC**		**Synth. Act. Score**		**CLIP Score**	
	Masks	Heatmaps	Masks	Heatmaps	Masks	Heatmaps	Masks	Heatmaps
Steering	<u>0.204</u>		0.773		1.473		0.182	
Top-k	0.194	0.186	0.782	0.857	1.453	1.609	0.188	0.187
Steering-informed Top-k	0.196	0.183	0.810	0.908	<u>1.691</u>	<u>2.156</u>	<u>0.190</u>	0.186

B. Later Layer Results

See Table 3.

C. Statistical Test Details

To assess statistical significance across explanation methods, we conduct pairwise one-tailed tests for each evaluation metric and masking type. Since evaluation scores are not normally distributed, as verified via a Shapiro-Wilk test, we apply the nonparametric Mann-Whitney U test. An explanation method is considered statistically significant if it is stochastically greater than both alternatives (with $p < 0.05$).

D. *Top-k* Explanation Evaluation Scores as a Function of Semantic Similarity Between *Steering* and *Top-k* Explanations

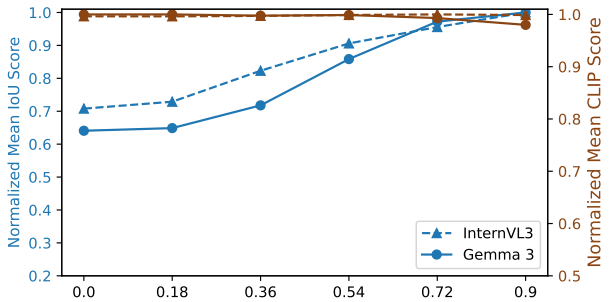

Figure 9. IoU Score and CLIP score values for *Top-k* method as a function of the similarity with *Steering* explanations.

E. Gemma 3 IoU and CLIP scores of *Steering* method as a function of the size of the LM m_{subj}

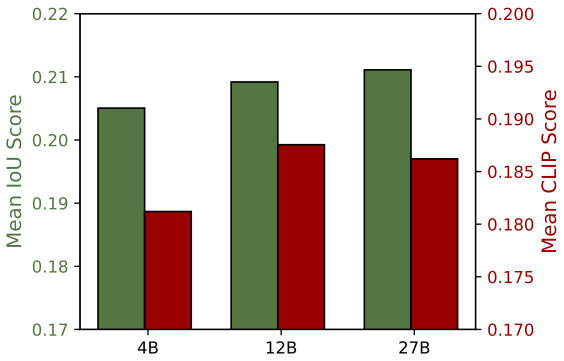

Figure 10. IoU Score and CLIP score values as a function of dataset size, for Masks Top-k method.

F. FLOPs Estimation

We compare the approximate Floating Point Operations (FLOPs) for generating explanations, using the estimate $2 \times$ Parameters $\times$ Tokens for a model forward pass [25]. Let:

- $N_{\text{eval}} = |\mathcal{D}^{\text{eval}}|$: size of the evaluation image set.
- P_{sub}: parameters of the subject model m_{sub} (also serving as m_{exp}).
- $P_{\text{SAE\_enc}} = d_{\text{model}} \cdot d_{\text{SAE}}$: parameters for the SAE.
- T_{img}: per image token representations (for m_{sub} input, for SAE processing per image, and for the empty image $\tilde{I}$. E.g., 4096 for Gemma 3).
- T_{prompt}: token count for the textual prompt.
- T_{expl}: max tokens in the explanation.
- k: number of top images selected.

***Top-k* Explanations.** This method consists of two main computational stages:

1. **Dataset Precomputation** (typically a one-time process to identify top-k activating images for features): It involves processing all N_{eval} images through m_{sub}, followed by SAE encoding for each representation using $\boldsymbol{W}_{\text{enc}}$. $\text{FLOPs}_{\text{precompute}} \approx N_{\text{eval}} \cdot T_{\text{img}} \cdot 2 \cdot (P_{\text{sub}} + P_{\text{SAE\_enc}})$. Aggregation and sorting costs are generally minor in comparison.
2. **Per-feature Explanation Generation**: The explainer model m_{sub} is conditioned on the prompt and the k selected images. $\text{FLOPs}_{\text{gen}} \approx 2 \cdot P_{\text{sub}} \cdot (T_{\text{prompt}} + k \cdot T_{\text{img}} + T_{\text{expl}})$.

The total cost is dominated by $\text{FLOPs}_{\text{precompute}}$ when N_{eval} is large.

***Steering*-based Explanations.** This approach avoids the dataset precomputation. An explanation for each feature i is generated via a single forward pass of m_{sub} from an intervention using the pre-defined SAE feature direction $\boldsymbol{W}_{\text{dec}}[i, :]$:

- **Per-feature Explanation Generation**: $\text{FLOPs}_{\text{steer}} \approx 2 \cdot P_{\text{sub}} \cdot (T_{\text{prompt}} + T_{\text{img}} + T_{\text{expl}})$. The costs for retrieving the SAE feature direction and applying the intervention (vector operations) are also incurred, in addition to the forward pass captured by the formula above.

***Steering-informed Top-k* Explanations.** This method combines the dataset precomputation with an intervened generation step:

1. **Dataset Precomputation**: This stage is identical to the corresponding stage in the *Top-k* method, incurring $\text{FLOPs}_{\text{precompute}}$ as defined above.
2. **Per-feature Explanation Generation**: Similar to standard *Top-k* generation, but with an intervention. The computational cost for generation remains approximated by $\text{FLOPs}_{\text{gen}}$ as defined for *Top-k* explanations. The costs for retrieving and applying the intervention are also incurred here, similar to the pure *Steering-based* method. This method achieves the best results at a comparable cost.

G. Prompts

G.1. Explainer Prompts

Steering Prompt

```
You are given an image highlighting a visual or semantic element. This element may
range from a low-level visual feature to a high-level abstract concept. Your task is to
describe this element in a single, clear sentence. If the element is a high-level
abstract concept, describe it as such; otherwise, describe its visual patterns.
Favor a more general interpretation. Start the highlighted element description
with \"The highlighted element in the image is a\".
```

Figure 11. Prompt used for obtaining explanations for the *Steering* method. Gemma 3 outputs when given this prompt and a blank image (without steering) are reported in Section G.2.

Top-k and Steering-informed Prompt (Masks)

```
You are given set of images highlighting a visual or semantic element. The patches of
the images not showing the element are masked out, giving the impression of a
pixelated image. This element may range from a low-level visual feature to a high-level
abstract concept. Your task is to describe this element in a single, clear sentence.
If the element is a high-level abstract concept, describe it as such; otherwise,
describe its visual patterns. Favor a more general interpretation. Provide a single
description for the highlighted element appearing in all images, and please ignore the
pixelated effect of the mask when describing the element. Start the highlighted element
description with \"The highlighted element in the image is a\".
```

Figure 12. Prompt used for obtaining explanations for the *Top-k* and *Steering-informed Top-k* method with Masks.

Top-k and Steering-informed Prompt (Heatmaps)

```
You are given set of images highlighting a visual or semantic element. The patches of
the images showing the element are highlighted with a green heatmap. This element may
range from a low-level visual feature to a high-level abstract concept. Your task is
to describe this element in a single, clear sentence. If the element is a high-level
abstract concept, describe it as such; otherwise, describe its visual patterns. Favor
a more general interpretation. Provide a single description for the highlighted element
appearing in all images, and please ignore the overlayed green heatmap when describing
the element. Start the highlighted element description with \"The highlighted element
in the image is a\".
```

Figure 13. Prompt used for obtaining explanations for the *Top-k* and *Steering-informed Top-k* method with Heatmaps.

G.2. Prompt-only behavior

To assess whether explanations may be driven by the prompt alone, we report model outputs when using the steering prompt with a blank image and no feature intervention. Smaller models correctly describe the absence of content (e.g., Gemma 3 4B: “solid, uniformly white space, creating a blank canvas effect”; Gemma 3 12B: “blank, white space”), while notably, the larger model produces a spurious description (Gemma 3 27B: “stylized depiction of a bird in flight, characterized by its curved wings and streamlined body”), which might be explained by the unnatural input.

H. Models and Datasets

We use the following assets in our work:

Models

Table 4. The list of models used in this work.

Model	Link	License
Gemma 3 [49]	Hugging Face (Google)	Gemma Terms of Use [9]
InternVL3-14B [56]	Hugging Face (OpenGVLab)	Apache 2.0
CLIP [38]	Hugging Face (OpenAI)	MIT License
SAM2 [41]	Hugging Face (Meta)	Apache 2.0
Stable Diffusion [12]	Hugging Face (Stability AI)	CreativeML OpenRAIL M license
all-mpnet-base-v2 [43]	HuggingFace	Apache 2.0

Datasets

Table 5. The list of datasets used in this work.

Dataset	Link	License
ImageNet [10]	Official Website	Custom (Non-commercial)

I. Compute Resources

All training and evaluation experiments were run on a single node of 4x NVIDIA Hopper H100 64GB GPUs. The demo website runs on a machine with 2x NVIDIA 4090 GPUs. Each Gemma 3 SAE training took approximately 6 hours on 1 GPU, and 3 hours for InternVL3.

J. Faithfulness Controls for Steering-Based Explanations

Table 6. Faithfulness controls for steering-based explanations on 1,000 SAE features (Gemma 3 vision encoder middle-layer). We compare true feature steering against random norm-matched directions and permuted feature directions. Mean scores shown. For Top-k-based explanations we used Masks.

Model	Explanation Method	IoU Score	AUROC	Synth. Act. Score	CLIP Score
Gemma 3	**Steering**	0.204	0.691	0.353	0.186
	Top-k	0.206	0.740	0.365	0.190
	Steering-informed Top-k	0.209	0.806	0.518	0.193
	Random-vector Steering	0.149	0.439	0.008	0.179
	Random-perm Steering	0.138	0.448	0.008	0.161

To assess whether steering-based explanations are feature-specific rather than driven by the prompt or generic activation effects, we introduce two control baselines: (i) random norm-matched directions, and (ii) random permutations of SAE feature directions. In both cases, the intervention strength is matched to the original steering setup.

We evaluate these controls on 1,000 SAE features from the middle layer of the Gemma 3 vision encoder, using Gemma 3 27B as the explainer language model. Table 6 reports the results.

We observe that both control baselines perform substantially worse than true feature steering across all metrics. In particular, synthetic activation drops to near-zero and AUROC converges to near 0.5 (random-classifier level), lack causal alignment with the underlying feature activations. IoU and CLIP scores are also consistently lower.