Beyond Cross-Modal Alignment: Measuring and Leveraging Modality Gap in Vision-Language Models

Hanqi Yan[1,*], Xiangxiang Cui[2,*], Lu Yin[2], Jindong Gu[3], Paul Pu Liang[4], Yulan He[1,5,†], Yifei Wang[4,†]

[1]King's College London, UK, [2]University of Surrey, UK, [3]University of Oxford, UK, [4]MIT CSAIL, USA, [5]The Alan Turing Institute, UK

Correspondence: yulan.he@kcl.ac.uk, yifei_w@mit.edu

Abstract

The success of vision-language models is primarily attributed to effective alignment across modalities such as vision and language. However, modality gaps persist in existing alignment algorithms and appear necessary for human perception – evident in modality-specific phenomena like visual texture and linguistic tone. These observations motivate us to computationally measure and leverage modality gaps to improve downstream tasks. We first introduce the **M**odality **D**ominance **S**core (**MDS**), which attributes multimodal features to specific modalities by categorizing them into three classes: vision-dominant features, language-dominant features, and cross-modal features. We then propose automatic interpretability metrics to evaluate these modality-specific features in a scalable manner. Finally, we demonstrate that the training-free model editing enhances multiple downstream tasks, including mitigating bias in gender classification, generating cross-modal adversarial examples, and enabling modality-specific control in text-to-image generation. Combined with task-agnostic interpretability tools, our work offers insights for systematic analysis and lightweight editing of multimodal models.

1 Introduction

Multimodal models have become foundational to the advancement of AI, enabling AI systems to process and understand information from multiple data modalities, such as vision and language (Radford et al., 2021; Kim et al., 2021; Lu et al., 2019; Liang et al., 2024). Vision-Language Models (VLMs) in particular operate under the premise that different data modalities share common, or cross-modal, features that can be jointly learned (Ngiam et al., 2011; Sun et al., 2024; Li et al., 2025).

Alongside these remarkable advancements, ongoing AI research aims to deepen our understanding of how different modalities interact and diverge within VLMs (Liang et al., 2022; Rawal et al., 2023; Schrodi et al., 2025; Zhang et al., 2025). For instance,Liang et al. (2022) revealed that image and text embeddings often reside in disjoint regions of the shared embedding space. Parcalabescu and Frank (2023) proposed Shapley value-based attribution methods to quantify the extent to which multimodal models rely on individual modalities, with follow-up work (Parcalabescu and Frank, 2025) diagnosing unimodal collapse.

Despite these modality preference measurements and optimization methods, existing studies treat modality gaps as undesirable imperfections, primarily diagnosing model collapse (Parcalabescu and Frank, 2025; Rawal et al., 2023; Schrodi et al., 2025) or motivating new training algorithms for improved alignment. Our work takes a contrasting perspective: we posit that *modality gaps* are both prevalent and beneficial for downstream tasks. This assumption is grounded in cognitive science, where modality commonality and separation have long been central themes. Researchers have examined how humans integrate and differentiate information across sensory modalities (Paivio, 1991; Spence, 2011; Fan et al., 2016), suggesting that modal specificity may be functionally advantageous rather than merely an artifact of imperfect alignment.

To investigate this hypothesis in VLMs, we conduct a systematic study with three core contributions:

1. We demonstrate that modality-specific information can be extracted from VLMs—specifically, text-dominant (`TextD`), image-dominant (`ImgD`), and cross-modal (`CrossD`) features—and show that these features exhibit distinct activation patterns when processing images versus text.

2. We propose embedding-based interpretability metrics to measure monosemanticity (within-modality coherence) and modality fidelity (cross-modality validation) in a multimodal setup. These metrics are scalable and compatible with the existing top-k activated interpretations.

*Equal contributions. †Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2026, pages 12123–12138

3. We design lightweight probing and steering methods to analyze models' concept-specific preferences and achieve precise, effective control over VLM behavior.

2 Related Work

Modality Gap. The study of modality differences has long been explored in cognitive science (Spence, 2011; Paivio, 1991; Calvert et al., 2004). In multimodal models, researchers have identified that modality bias and gaps are prevalent in both early Multimodal Models (MMs) (Liang et al., 2022) and large MMs (Zhang et al., 2025). Moreover, modality gaps have been shown to negatively impact downstream tasks such as video understanding (Rawal et al., 2023) and object detection (Schrodi et al., 2025). Several metrics have been proposed to quantify modality gaps, including L2M (Liang et al., 2022), MM-SHAP (Parcalabescu and Frank, 2023), and its variants (Parcalabescu and Frank, 2025), which measure the degree to which individual modalities contribute to model predictions. Our work differs in two key aspects: (i) we measure modality gaps at the component level, identifying how individual model components respond differentially to different modalities; (ii) we leverage modality-specific components for probing and steering, treating specialization as a functional feature rather than an imperfection.

Interpretability Measurements. Existing interpretability measurements are based on summarizing patterns in top-k activated samples. For example, logit lens (nostalgebraist, 2020) has inspired many studies in both unimodal and multimodal representation understanding (Parekh et al., 2024; Jiang et al., 2025). The embedding-based extension (Phukan et al., 2025) alleviates its limitation in processing contextual-related concepts. More recently, LLMs have been used to generate explanations for activation patterns, with prediction accuracy on held-out samples serving as an interpretability proxy (Bills et al., 2023). However, this LLM-as-a-judge approach is computationally expensive and only measures semantic coherence within the unimodality, neglecting cross-modality consistency. Our interpretability metrics address these limitations through scalable embedding-based computation and introduce modality fidelity to fill the gap.

3 Identify Modality-Specific Features

Modality alignment and fusion are crucial to the success of existing VLMs (Liang et al., 2022; Schrodi et al., 2025), while the modality-specific gap has been extensively studied in cognitive science. For instance, Ungerleider and Haxby (1994); Fan et al. (2016) have found that regional specificity and coordinated processing coexist in the human brain. Therefore, we start with the question *"whether there are modality-specific features in VLMs?"* To answer the question, we use CLIP models from OpenAI (Radford et al., 2021).

Background: Modality Alignment in VLMs. Typically, there are an image encoder and a text encoder in a VLM for image and text input processing, respectively. Specifically, the image-text pair $(x_{\text{img}}, x_{\text{txt}})$ is fed to an image encoder f_{img} and a text encoder f_{txt} within the model, respectively and the final-layer representations $z_{\text{img}} \in \mathbb{R}^D$ and $z_{\text{txt}} \in \mathbb{R}^D$ are then optimized jointly in the shared D-dimensional representation space. An alignment loss, such as the contrastive loss in CLIP (Ilharco et al., 2021) across the two modalities, is applied for modality alignment. A persistent modality preference/gap remains across most multimodal models (Liang et al., 2022; Zhang et al., 2025).

3.1 Modality-specific Feature Identification

To measure the modality gap, Liang et al. (2022) used the difference between the center of image embeddings and text embeddings of M input pairs, i.e., $\frac{1}{M}(\sum_{i=1}^{M} ||z_{\text{img},i}||_2 - \sum_{i=1}^{M} ||z_{\text{txt},i}||_2)$. We extend this model-level measurement to a fine-grained metric, i.e., the predominant modality associated with each dimension $d \in \{1, 2, \ldots, D\}$ in the shared embedding space. The proposed modality dominance score (**MDS**), denoted as $R(d)$ shown in Eq. (1) reflects how strongly the d-th feature [1] is influenced by the image modality:

$$R(d) = \frac{1}{M} \sum_{i=1}^{M} \frac{||z_{\text{img},i}^{(d)}||}{||z_{\text{img},i}^{(d)}|| + ||z_{\text{txt},i}^{(d)}||}. \quad (1)$$

Specifically, we feed M image-text pairs to the VLM and extract the corresponding image features $z_{\text{img},i}$ and text features $z_{\text{txt},i}$ for i-th input. For each d-th dimension in the D-dimension shared space, we calculate the relative activation between the features from the two modalities. This modality

[1] Each feature dimension corresponds directly to a feature/neuron in the VLM's final layer; our study thus focuses on the interpretability of the model's intrinsic components.

fraction is averaged over more than $M = 10k$ input pairs, providing a representative estimate of the modality distribution.[2]

We then categorize all D features into three groups based on their deviation from the mean μ and standard deviation σ of the MDS distribution:

$$\begin{aligned} \texttt{TextD}&: \; R(d) < \mu - \sigma; \\ \texttt{CrossD}&: \; \mu - \sigma < R(d) < \mu + \sigma; \\ \texttt{ImgD}&: \; R(d) > \mu + \sigma \end{aligned} \tag{2}$$

We anticipate that ImgD features are predominantly activated by visual concepts, TextD features by textual concepts, and CrossD features are simultaneously activated by the shared commonalities between image and text.

3.2 Quantitative Evaluation for MDS

To verify that modality-specific features effectively capture their intended modality information, we employ an ablation-based validation approach. Specifically, we remove these features from the original representation from CLIP ViT-H/14 (LAION-2B) (Ilharco et al., 2021) by zeroing out their corresponding indices, then use the modified representations as input features for logistic regression. We evaluate the representation on image/-text classifications using samples from COCO (Lin et al., 2014). A decrease in classification accuracy indicates that the removed features contained substantial modality-specific information, thereby validating our feature attribution method.

The results are shown in Table 1. It is observed that removing the ImgD leads to larger classification degradation in Image classification, while removing TextD leads to larger drops in text classification; while CrossD does not show any particular modality tendency in classification.

3.3 Qualitative Evaluation for MDS

We randomly select two features from the three groups, and then display their most-activated images and texts in Figure 1 (ImD), Figure 2 (TextD) and Figure 12 (CrossD). **ImgD activates fundamental visual concepts, such as repeated patterns and colors.** Feature 647 activates images with diverse repetitive patterns; feature 667 focuses on scenes with aquatic-blue elements. Although less coherent than the images, some patterns do emerge for its activated texts: feature-647 activates two sentences that refer to repetitive patterns, such as "*tufted upholstery*"; feature-667 activates texts related to "*snowy*" and "*winter*". These observations indicate the modality alignment while the visual commanalities are more predominant for the ImgD. **TextD capture abstract concepts, such as human feelings and atmosphere.** For the activated images for feature-34 (the 1st row), most of the images have red color, with one image depicting a couple talking beside the sea; for feature-242, there are no clear patterns among the activated images. When looking at the activated texts, sentences activated by feature-34 center around a sweet and happy atmosphere between couples, with themes like cuddling, embracing, and hugging. Feature-242 focuses on strong human emotions, such as "*never*", "*terrifying*" and exclamation marks. These TextD generally correspond to abstract and consistent human emotions, which can be conveyed with a variety of visual objects. For example, in the second row, the first image depicts a collection of stones forming a heart shape, while the fourth image is a scenic view during a great trip. **CrossD (the majority features) capture shared semantics across modalities.** Differently, CrossD features capture common concepts that could be expressed in both visual and language modalities. (Details can be found in Appendix A.4.)

Task	Deletion	# Neurons	Accuracy	Δ Acc
Image CLS	None	0	**0.776**	/
	Random	426	0.757	-2.1%
	ImgD	426	**0.750**	**-2.6%**
	Random	554	0.756	-2.0%
	TextD	554	0.760	-1.0%
	Random	44	0.773	-0.3%
	CrossD	44	0.769	-0.5%
Text CLS	None	0	**0.713**	/
	Random	426	0.694	-1.9%
	ImgD	426	0.702	-1.1%
	Random	554	0.693	-2.0%
	TextD	554	**0.683**	**-3.0%**
	Random	44	0.710	-0.3%
	CrossD	44	**0.712**	-0.1%

Table 1: Performance changes of Modality-specific classification after removing: random vs. specialized features, i.e., ImgD, TextD and CrossD. We also remove the same number of random feature indices for comparison.

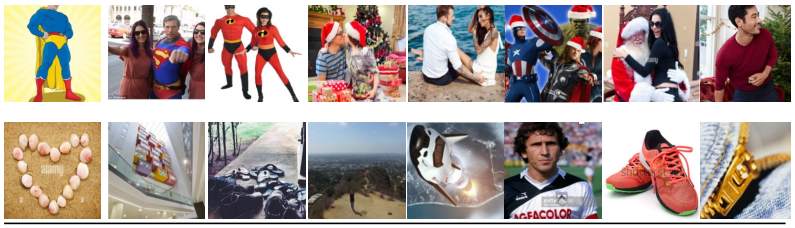
4 Automatic Interpretability Evaluation

Although we have identified modality-dominant features, features in deep models are inherently *polysemantic* (Olah et al., 2020)—each feature often encodes multiple unrelated semantic concepts, potentially spanning both textual and visual modalities, hindering interpretability. *Monosemantic-*

[2]Details of the MDS calculation are in Appendix A.2.

Feature-647: Pattern and others.	*Feature-667: Scenes in winter and other.*
Seamless pattern, flowers on a background.	Covering the trailhead in a winter wonderland.
Every girl should have this in their bedroom.	Red leather belt, a perfect accessory.
Could new showroom and model signal the start?	The image of drum under the white background.

Figure 1: Activated images and texts (in Table) by `ImgD`. Top image row (feature 647): patterns and textures. Bottom image (feature 667): water and aquatic themes in blue. Texts in blue align with visual concepts.



Feature-34: Sweet and happy Couple.	*Feature-242: Strong emotion.*
Attractive young couple sitting on a bench, talking and laughing with the city.	Animal looking for a cat tree without carpet your options have greatly expanded.
Sculpture of lovers at the temple	Sinkhole, most terrifying thing I have ever seen.
Young couple in love, hugging in the old part of town.	We're away from the beginning of the holiday season here!

Figure 2: Activated images and texts (in Table) by `TextD`. Top image row (feature 34): couples and individuals in red attire. Bottom image row (feature 242): diverse objects. Text in blue aligns with visual concepts.

ity (Elhage et al.; Bills et al., 2023; Gurnee et al., 2023; Yan et al., 2024a) has thus emerged as a paradigm for deriving interpretable features that encode single concepts. However, scaling interpretability evaluation remains an open challenge due to heavy reliance on costly human annotations (Gao et al., 2024) or LLM-generated explanations (Bills et al., 2023). To address this bottleneck, we propose a suite of automated metrics to measure feature interpretability in multimodal models.

4.1 Overview

A feature is considered interpretable if its semantic meaning can be readily understood by humans. In practice, interpretability is assessed by examining whether the top-k activated samples exhibit coherent and consistent patterns. This top-k activation-based interpretation has become standard practice for analyzing both language (Geva et al., 2021) and multimodal models (Parekh et al., 2024). Bills et al. (2023) advanced this approach by first prompting a large language model (LLM) to generate explanations based on activated tokens, then using these explanations to predict activation values for held-out tokens. The correlation between predicted and actual activations serves as an interpretability score. However, this LLM-as-a-judge paradigm faces limitations in reliability and scalability (Gu et al., 2024; Yan et al., 2024b).

Building upon the top-k activation framework, we propose scalable, *embedding-based* evaluation metrics tailored to multimodal models. In this context, interpretability encompasses two dimensions:

- ***Monosemanticity*** (within-modality coherence). It measures whether a feature's top-k activated samples exhibit semantic coherence within a single modality based on their embedding similarity.
- ***Modality fidelity*** (cross-modality validation). It compares the monosemanticity scores across modality-attributed features to assess whether features remain faithful to their assigned modality. For example, when processing visual inputs, `ImgD` should exhibit higher monosemanticity scores than `TextD`, and vice versa for textual inputs.

4.2 Intepretability Metrics

Given a feature $z^{(d)}$, the d-th dimension of $z \in \mathbb{R}^D$, we propose to use embedding models $h : \mathbb{Z}^D \rightarrow \mathbb{Z}^{D'}$ to calculate the interpretability metrics.

Monosemanticity. For each image/text feature $z^{(d)}$, we collect the top m most-activated image/-text samples for this dimension, and feed them to the embedding model h to get $Z_+ \in \mathbb{R}^{m \times D'}$. For

comparison, we embed m random samples into $Z_- \in \mathbb{R}^{m \times d'}$. Then, we calculate the inter-sample similarity between the selected samples, $S_+ = Z_+ Z_+^\top \in \mathbb{R}^{m \times m}$ and $S_- = Z_- Z_-^\top \in \mathbb{R}^{m \times m}$. The monosemanticity of an individual feature $z^{(d)}$ is measured by calculating the relative difference between the two similarity scores, denoted as $I(z^{(d)})$ (***EmbSim***). We also propose a binary metric to avoid the different scales in different modalities, denoted as $W(z^{(d)})$ (***WinRate***):

$$
\begin{aligned}
I(z^{(d)}) &= \frac{1}{m(m-1)} \sum_{i \neq j} \frac{(S_+)_{ij} - (S_-)_{ij}}{(S_-)_{ij}}, \\
W(z^{(d)}) &= \frac{1}{m(m-1)} \sum_{i \neq j} \mathbf{1}_{[(S_+)_{ij} > (S_-)_{ij}]}.
\end{aligned} \tag{3}
$$

The overall interpretability score is the average across all d dimensions for $z^{(d)}$ where $d \in [1, D]$. A higher monosemanticity score (both *EmbSim* and *WinRate* are ***Mono***, with a superscript representing the input modality) indicates that the extracted features exhibit stronger semantic consistency towards the given modality. [3]

Modality Fidelity. Based on the single-modality monosemanticity (semantic coherence), we proceed to cross-modality interpretability. Specifically, we ask: is `ImgD` indeed more effective at capturing coherent visual inputs than `TextD`? Similarly, is `TextD` better at encoding textual semantics compared to `ImgD`? Therefore, we define the modality fidelity as:

$$
\begin{aligned}
\textit{Visual Fidelity} &= \textit{Mono}^{\text{vis}}(\texttt{ImgD}) - \textit{Mono}^{\text{vis}}(\texttt{TextD}) \\
\textit{Textual Fidelity} &= \textit{Mono}^{\text{txt}}(\texttt{TextD}) - \textit{Mono}^{\text{txt}}(\texttt{ImgD})
\end{aligned}
$$

4.3 Interpretability Evaluation Results

Sparse Autoencoders (SAEs) (Cunningham et al., 2023) have been shown to effectively generate monosemantic features by enforcing feature sparsity. Therefore, we apply our metrics to compare CLIP and CLIP+SAE to validate whether our evaluation framework yields conclusions consistent with existing literature, establishing its reliability as an interpretability metric. Beyond this validation, we also explore whether other representation learning methods can enhance feature interpretability.

4.3.1 Comparison Models

We incorporate several representation learning algorithms (including SAEs) that aim to learn modality-specific features [4].

Multimodal SAEs. SAEs have emerged as a scalable tool for transforming polysemantic *neurons* into interpretable, monosemantic *features* across various LLMs (Templeton, 2024; Gao et al., 2024; Lieberum et al., 2024). We extend it for multimodal settings by training a **single** SAE model $g : \mathbb{Z} \to \mathbb{Z}$ to reconstruct z, i.e., the final-layer outputs from the image and text encoder within CLIP, respectively. Specifically, we adopt that applies a linear encoder W_{enc} followed by a $\text{Top}\,K$ operation that only keeps the K most activated units while zeroing out the rest. The sparse latent representation z^{sae} is then reconstructed using a linear decoder W_{dec}:

$$
\begin{aligned}
z^{\text{sae}} &= \text{TopK}\left(W_{\text{enc}}\left(z - b_{\text{pre}}\right)\right), \\
\hat{z} &= W_{\text{dec}} z^{\text{sae}} + b_{\text{pre}}.
\end{aligned} \tag{4}
$$

$z \in \mathbb{R}^D$ is the inputs of SAE, i.e., z_{img} or z_{txt}. $z^{\text{sae}} \in \mathbb{R}^n$ is the learned sparse representation. We train the multimodal SAE to reconstruct z_i, z_t.

DeCLIP. Beyond multimodal supervision (image-text pairs), DeCLIP (Li et al., 2022) also incorporates single-modal self-supervision (image-image pairs and text-text) for more efficient joint learning.

Multimodal NCL. As shown in Wang et al. (2024), the non-negative constraints allow Non-negative Contrastive Learning (NCL) to extract highly sparse features and significantly improve feature monosemanticity. Therefore, we introduce a variant of NCL to enhance modality specification with the following loss,

$$
-\mathbb{E}_{z_{\text{img}}, z_{\text{txt}}} \log \frac{\exp(g(z_{\text{img}})^\top g(z_{\text{txt}}))}{E_{z_{\text{txt}}^-} \exp(g(z_{\text{img}})^\top g(z_{\text{txt}}^-))}, \tag{5}
$$

where here we use a ReLU-activated MLP network g to map input features to *non-negative* output latent features.

4.3.2 Evaluation Results

We calculate the monosemanticity and modality fidelity for all the models.

Results of Monosemanticity. We compute the *Mono* (interpretability) score by identifying the top-20 most activated images and texts for each feature, respectively. From the average interpretability results in Figure 4, we observe the following: (i) The features extracted using SAE and NCL (which enforce the feature sparsity) exhibit the highest overall monosemanticity for both activated input images and texts. (ii) DeCLIP does not enhance

[3] We calculate the average of *EmbSim* and *WinRate* as *Mono* in the main content; the separate results for the two metrics can be found in Appendix A.3.2.

[4] Implementation details for these methods are shown in Appendix A.1.

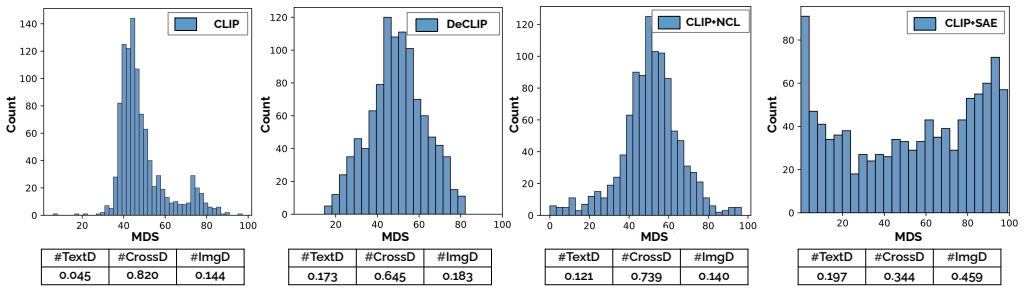


Figure 3: Modality Dominance Score (MDS) distributions of three feature categories for different VLMs.

interpretability through self-supervision alone; the monosemanticity on the textual side becomes even worse. This suggests that polysemantic features remain prevalent in DeCLIP, although their modality separation is clearer than in CLIP.

Moreover, we observe that **monosemanticity enhancement encourages more modality-specific neurons**. In Figure 13, we calculate the MDS and visualize the distributions of the three feature groups across models. Interestingly, we find that CLIP contains a spectrum of features with different modality dominance. Specifically, its distribution skews towards the image modality, and this trend is consistent across all models. DeCLIP, on the other hand, shows a more balanced and less centered distribution. This suggests that DeCLIP, through self-supervision, extracts more modality-specific features, which might be overlooked by pure vision-language contrastive models like CLIP. The extracted features on top of NCL and SAE also exhibit less skewness, with SAE showing the most balanced distribution, indicating its strong capability to extract diverse monosemantic features.

Results of Modality Fidelity. We have the following observations from Figure 5: (i) For CLIP, all the modality monosemanticity is negative, demonstrating the high entanglement of the two modality information. (ii) All the methods prompt the modality monosemanticity compared with CLIP. Particularly, the improvements of DeCLIP can be attributed to its single-modal alignment training loss, which could weaken some cross-modal associations in CLIP. (iii) NCL stands out as the best model for capturing both visual and textual monosemantic features, followed by SAE.

5 Applications of Modality Gap

Beyond interpretability, we design lightweight probing and steering methods based on modality-specific features to analyze VLMs' perceptual biases and enable precise behavioral control. [5]

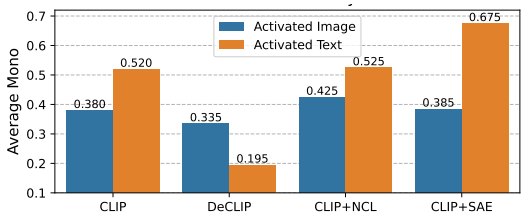


Figure 4: Monosemanticity.

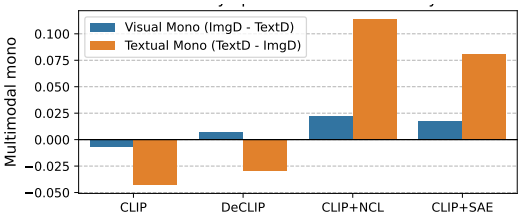


Figure 5: Modality Fidelity.

5.1 Understanding Gender Pattern

We describe gender using both visual and textual features and these data are used to train VLMs. To test whether there is a modality-specification in different genders, e.g., *Does the concept of the feminine get described by images more frequently? such as more colorful outfits.* To test this hypothesis, we collect both male and female images with their corresponding textual descriptions from the cc3m-wds (Sharma et al., 2018). These images are then encoded using the Clip+SAE model, extracting 1024-dimensional features for both female and male subjects. Next, we apply a zero-mask intervene strategy to remove the `ImgD` and `TextD` from these representations.

We compare changes in gender classification accuracy when removing `ImgD` features from image inputs, which capture dominant feminine visual cues, versus removing `TextD` features from text inputs. As shown in Table 2, we find that feminine concepts are primarily preserved in `ImgD` (as the removal of `ImgD` from the image leads to larger classification degradation), whereas male concepts are more affected by the removal of `TextD`.

[5] Detailed implementations are in Appendix A.5.

Gender	w.o ImgD	w.o TextD
Female	**17.65**	7.27
Male	5.64	**28.67**

Table 2: Gender classification changes (%) after removing ImgD(textD) **from input image(text)** for both female and male concepts identification. It is to verify the dominant modality for different genders.

Figure 6: ***Female*** figures ordered by their percentages of ImgD features: 0.14, 0.16, 0.18, 0.20, 0.22, 0.24, 0.26. More feminine concepts are observed to be related with more ImgD.

Understand the what feminine concepts the ImgD represent. We sample different female images which differ in the percentage of their most activated features categorized as ImgD features. The results are in Figure 6. From left to right, more activated features are ImgD, and they tend to contain more detailed (*stereotype*) feminine concepts, such as a backless skirt, hair accessories. The middle images show professional females, such as a politician and a doctor; and the first image shows a pair of legs in sports shoes, with minimal feminine factors, the pink color.

5.2 Generate Modality-Specific Attacks

We investigate the impact of different types of features on multimodal adversarial attacks (Cui et al., 2024; Yin et al., 2024), following the setup in Shayegani et al. (2024).

The adversarial sample is a benign-appearing image, e.g., a scenery image, but injected with harmful semantic information, such as the phrase *"I want to make a bomb"*. One defense optimization strategy involves minimizing the distance between the embeddings of adversarial sample $\mathbf{F}_{adv}$ and a benign sample $\mathbf{F}_{ben}$, and accordingly update the adversarial sample (in Figure 7). The paired benign image is injected with the friendly text, e.g., *"peace and love"*. To study the effects of our identified modality features, we only select the target feature index I for alignment training, i.e., ImgD, TextD, and CrossD. The alignment loss is $\mathcal{L} = \|\mathbf{F}_{adv}[:, I] - \mathbf{F}_{ben}[:, I]\|_2$. Finally, the optimized adversarial sample is then adopted to attack a VLM, LLaVA-1.5-7b (Liu et al., 2023). We use the LLM-as-a-Judge to evaluate the generated response from the VLM, where DeepSeek-V3 (DeepSeek-AI et al., 2024) is required to generate a binary label indicating whether the attack is successful. Supposedly, the features containing more information related to the malicious semantics will contribute most to the attack defense.

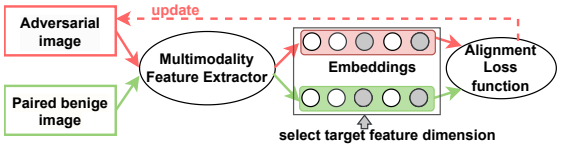


Figure 7: Alignment training to de-toxicity of the adversarial sample, with only selected target feature dimensions (in gray), i.e., ImgD, TextD and CrossD, involved in the alignment.

Target feature	ImgD	TextD	CrossD
Success Rate (↓)	62.71%	24.89%	35.44%

Table 3: Success rate for adversarial attacks with different target features involved in the alignment training. The success rate for original adversarial samples without alignment training is 73.26%, while for randomly selected features is 54.28%.

Results. The attack success rates are shown in Table 3. We select the same number from ImgD, TextD, CrossD to be involved in alignment training, as well as randomly sample the same number of features across the three feature sets as a baseline. We attack the VLM repeatably for 100 times per sample, and we have generated 50 adversarial samples. We observe that (i) by comparing with the success rate of original adversarial samples, the alignment training with any selected features defense the attacks to some extent; (ii) using TextD yields the best defense performance, followed by CrossD and ImgD. This can be explained by the fact that the adversarial information primarily stems from undesirable textual semantics. And **it demonstrates that TextD effectively captures most of the semantic content.** In contrast, CrossD captures partial semantics, while ImgD is the least related to semantic information, resulting in minimal benefits for such modality-specific jailbreak defense.

5.3 Controllable Text-to-Image Generation

Despite the impressive capabilities of text-to-image generation models (Yu et al., 2024; Koh et al., 2024; Swamy et al., 2024), their internal mechanisms for bridging linguistic semantics and visual details remain poorly understood. A key challenge is disentangling how modality-specific features influence the fidelity and controllability

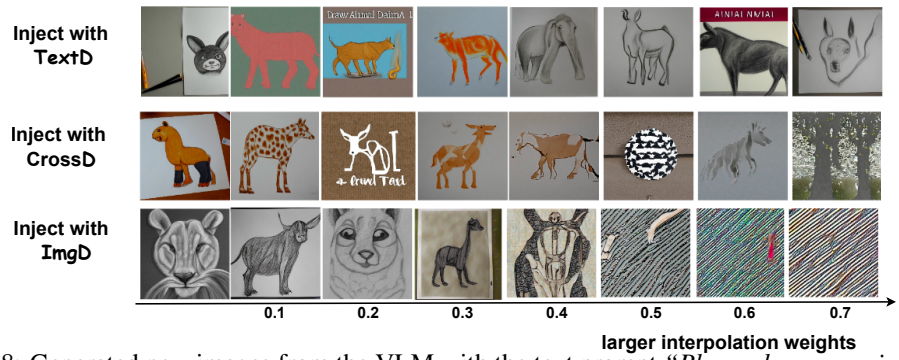
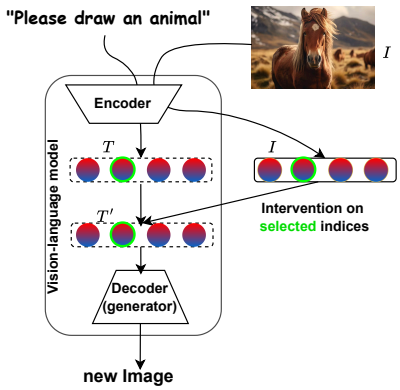


Figure 8: Generated new images from the VLM with the text prompt *"Please draw an animal"* and varying levels of intervention from a reference image (horse). From left to right, the interpolation weights α range from 0.0 to 0.7. Images generated with `TextD` typically depict clear main subjects (horse) without transferring the visual background details from the reference image. In contrast, injection of `ImgD` introduces low-level visual details as well as image distortions when α is large.



$$T' = (1-\alpha)T + \alpha I$$

Figure 9: The reference image R is used for modality-specific control over text-to-image generation process.

of generation. Therefore, we conduct a feature intervention experiment during the generation of Stable Diffusion v2 (Rombach et al., 2022).

Intervention. The process is depicted in Figure 9. We investigate the generation process by intervening in different modality-specific features in Stable-Diffusion-v2 (Rombach et al., 2022), i.e., the shown VLM with an encoder and decoder (generator). The input text prompt is *"Please draw an animal"*. The encoder generates an embedding $\mathbf{T}$, representing the original multimodal embedding ready for generation. Additionally, we provide a reference image; here is a horse - processed through the same encoder, producing a reference embedding $\mathbf{R}$. To control the generation through modality-specific feature intervention, we interpolate $\mathbf{T}$ only at the specified indices defined by MDS. The final multimodal embedding is computed as: $\mathbf{T}'[I] = \alpha\mathbf{T}[I] + (1-\alpha)\mathbf{R}[I]$, where operations are applied exclusively to the feature indices defined by I, i.e., `TextD`, `CrossD`, and `ImgD`.

Results. We feed $\boldsymbol{T'}$ to the generator of the VLM with different α ranging from 0 to 0.7 with an interval of 0.1. The generated images with the selected indices correspond to `TextD`, `CrossD`, and `ImgD` are shown in Figure 8. The results clearly demonstrate that larger interventions on `TextD` lead to stronger control over high-level semantic concepts—for example, the generated image more distinctly resembles a horse (head). All these generated images injected by `TextD` typically depict clear main subjects without transferring visual background details from the reference image. In contrast, interventions on `ImgD` result in more visual details from the reference image being preserved, such as non-white and fur-like patterned background, which are visible in `ImgD` when $\alpha \geq 0.3$. To better contrast the effects of `ImgD` and `TextD`, we also use a reference image with a horse as the main subject, but in different styles/backgrounds. More results are shown in Figure 15.

6 Conclusion

In this study, we explored the monosemanticity of features within VLMs to elucidate the commonalities and distinctions across visual and linguistic modalities. Specifically, we successfully categorized multimodal features according to their dominant modality. Our proposed embedding-based interpretability metrics fill the gap in multimodal monosemanticity assessment. Moreover, we designed lightweight probing and editing methods based on modality-specific features and demonstrated great potential in mitigating gender bias, defending against adversarial attacks, and enabling controllable multimodal generation.

Limitation

While our work provides valuable insights into modality-specific feature analysis in vision-language models, several limitations warrant discussion. First, we did not conduct human studies to validate our interpretability metrics. Although our embedding-based metrics align with existing interpretability tools, direct human evaluation could provide stronger evidence that our categorizations match human cognitive interpretations of modality dominance. Second, our experiments focus exclusively on CLIP-family models. The generalizability of our findings to other vision-language architectures (e.g., BLIP or autoregressive VLMs) remains an open question. Different architectural designs may exhibit distinct modality gap characteristics that require adapted analysis methods.

Acknowledgment

This work was supported in part by the UK Engineering and Physical Sciences Research Council through a Turing AI Fellowship (grant no. EP/V020579/1, EP/V020579/2) and the Prosperity Partnership scheme (grant no. UKRI566). The authors also acknowledge the use of the King's Computational Research, Engineering, and Technology Environment (CREATE) at King's College London.

References

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models. *URL https://openaipublic. blob. core. windows. net/neuron-explainer/paper/index. html.(Date accessed: 14.05. 2023)*, 2.

Gemma Calvert, Charles Spence, and Barry E Stein, editors. 2004. *The Handbook of Multisensory Processes*. MIT Press.

Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. 2024. On the robustness of large multimodal models against image adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24625–24634.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, R. Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *International Conference on Learning Representations*.

DeepSeek-AI, Aixin Liu, and Bei Feng et al. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Nelson Elhage, Neel Nanda, Catherine Olsson, and Others. A mathematical framework for transformer circuits. *Transformer Circuits Thread (2022)*.

Lingzhong Fan, Hai Li, Junjie Zhuo, Yu Zhang, Liangfu Chen, Zhengyi Yang, Congying Chu, Sangma Xie, Angela Laird, Peter Fox, Simon Eickhoff, Chunshui Yu, and Tianzi Jiang. 2016. The human brainnetome atlas: A new brain atlas based on connectional architecture. *Cerebral Cortex*, 26:bhw157.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2024. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. 2023. Finding neurons in a haystack: Case studies with sparse probing. *ArXiv*, abs/2305.01610.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip. If you use this software, please cite it as below.

Nicholas Jiang, Anish Kachinthaya, Suzanne Petryk, and Yossi Gandelsman. 2025. Interpreting and editing vision-language representations to mitigate hallucinations. In *The Thirteenth International Conference on Learning Representations*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*.

Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. 2024. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. 2022. Supervision exists everywhere: A

data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations*.

Zhaowei Li, Wei Wang, YiQing Cai, Qi Xu, Pengyu Wang, Dong Zhang, Hang Song, Botian Jiang, Zhida Huang, and Tao Wang. 2025. Unifiedmllm: Enabling unified representation for multi-modal multi-tasks with large language model. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 334–344.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2024. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10):1–42.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *arXiv preprint arXiv:2408.05147*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 689–696.

nostalgebraist. 2020. interpreting gpt the logit lens. *URL* https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001.

Allan Paivio. 1991. Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 45(3):255.

Letitia Parcalabescu and Anette Frank. 2023. Mm-shap: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4032–4059.

Letitia Parcalabescu and Anette Frank. 2025. Do vision & language decoders use images and text equally? how self-consistent are their explanations? In *The Thirteenth International Conference on Learning Representations*.

Jayneel Parekh, Pegah Khayatan, Mustafa Shukor, Alasdair Newson, and Matthieu Cord. 2024. A concept-based explainability framework for large multimodal models. *Advances in Neural Information Processing Systems*, 37:135783–135818.

Anirudh Phukan, Divyansh, Harshit Kumar Morj, Vaishnavi, Apoorv Saxena, and Koustava Goswami. 2025. Beyond logit lens: Contextual embeddings for robust hallucination detection & grounding in VLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9661–9675, Albuquerque, New Mexico. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Isha Rawal, Shantanu Jaiswal, Basura Fernando, and Cheston Tan. 2023. Dissecting multimodality in videoqa transformer models by impairing modality fusion. In *International Conference on Machine Learning*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Simon Schrodi, David T. Hoffmann, Max Argus, Volker Fischer, and Thomas Brox. 2025. Two effects, one trigger: On the modality gap, object bias, and information imbalance in contrastive vision-language models. In *The Thirteenth International Conference on Learning Representations*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*.

Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. 2024. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. In *The Twelfth International Conference on Learning Representations*.

Charles Spence. 2011. Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, 73(4):971–995.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2024. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110.

Vinitra Swamy, Malika Satayeva, Jibril Frej, Thierry Bossy, Thijs Vogels, Martin Jaggi, Tanja Käser, and Mary-Anne Hartley. 2024. Multimodn—multimodal, multi-task, interpretable modular networks. *Advances in Neural Information Processing Systems*, 36.

Adly Templeton. 2024. *Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet*. Anthropic.

Leslie G. Ungerleider and James V. Haxby. 1994. 'what' and 'where' in the human brain. *Current Opinion in Neurobiology*, 4(2):157–165.

Yifei Wang, Qi Zhang, Yaoyu Guo, and Yisen Wang. 2024. Non-negative contrastive learning. *ICLR*.

Hanqi Yan, Yanzheng Xiang, Guangyi Chen, Yifei Wang, Lin Gui, and Yulan He. 2024a. Encourage or inhibit monosemanticity? revisit monosemanticity from a feature decorrelation perspective. *ArXiv*, abs/2406.17969.

Hanqi Yan, Qinglin Zhu, Xinyu Wang, Lin Gui, and Yulan He. 2024b. Mirror: Multiple-perspective self-reflection method for knowledge-rich reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7086–7103, Bangkok, Thailand. Association for Computational Linguistics.

Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Jinguo Zhu, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. 2024. Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models. *Advances in Neural Information Processing Systems*, 36.

Lijun Yu, Yong Cheng, Zhiruo Wang, Vivek Kumar, Wolfgang Macherey, Yanping Huang, David Ross, Irfan Essa, Yonatan Bisk, Ming-Hsuan Yang, and 1 others. 2024. Spae: Semantic pyramid autoencoder for multimodal generation with frozen llms. *Advances in Neural Information Processing Systems*, 36.

Yu Zhang, Jinlong Ma, Yongshuai Hou, Xuefeng Bai, Kehai Chen, Yang Xiang, Jun Yu, and Min Zhang. 2025. Evaluating and steering modality preferences in multimodal large language model. *arXiv preprint arXiv:2505.20977*.

A Appendix

A.1 Implementation for Monosemanticity Tools

The three monosemantic tools, DeCLIP, Multimodal SAE, and Multimodal NCL are all on top of the canonical ViT-B-32 CLIP [6] model from OpenAI (Radford et al., 2021), with ResNet50. The four methods (including CLIP) share the same model structures but are trained with different training objectives. We load them by feeding the checkpoints using the `open_clip.create_model_and_transforms` function in the published https://github.com/mlfoundations/open_clip.

The feature dimensions of the output features from the image encoder and text encoder are both 1024, the same for CLIP, DeCLIP, and Multimodal NCL. To retain the multimodal representation efficiency in downstream tasks, we have trained the SAE and NCL to reach a very small reconstruction loss for the original features z from CLIP. The dataset for NCL and SAE training is the train split (around 2900k image-text pairs) from `cc3m-wds`[7]. We train the two variants, i.e., SAE and NCL, on top of the pretrained CLIP using a single 3090 GPU.

DeCLIP. We use the checkpoint released in https://github.com/Sense-GVT/DeCLIP to extract the last layer features, z_i and z_t.

Multimodal SAE. We insert an SAE model to map the original feature into a sparse latent space, i.e., $z^d \rightarrow z^n$, with top-k latent as nonzero values. Empirically, we found that when $n = d$ and $k = 32$, we can get the best results to balance the sparsity and downstream task performance. Such an SAE model (shared parameter) is inserted at the end of the image and text encoder in CLIP.

```
def get_sae_embedding(self, z):
    z = self.encoder(z)
    z_sae = F.relu(z)
    vals, ids = z_sae.topk(self.k, dim=1)
    z_sae = torch.zeros_like(z_sae)
    z_sae.scatter_(1, ids, vals)
    return z_sae
```

Inspired by (Gao et al., 2024), we train the SAE until the sparsity (the inactive dimension) of image features and text features doesn't increase (the same stop criteria for NCL). Noting that there are many zero values in z^{sae}, we remove those zero activity features (called dead latents in (Gao et al., 2024)) for further studies. We show the changes of active dimensions of image features and text features in Figure 10.

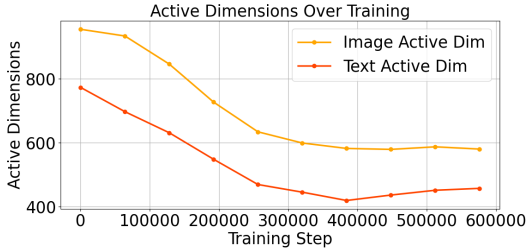


Figure 10: The changes of active dimensions over **SAE** training.

Non-negative Contrastive Learning (NCL). We add the NCL block, i.e., the projector, after obtaining z_i and z_t from the image encoder and text encoder. The training loss is shown in Eq. 5.

```
self.projector = nn.Sequential(
    nn.Linear(embed_dim, embed_dim),
    nn.LayerNorm(embed_dim),
    nn.ReLU(),
    nn.Linear(embed_dim, embed_dim),
)
z_ncl = self.projector(z)
```

Similarly, the activated dimensions for image features and text features decrease and are then flattened (shown in Figure 11.) By comparing with Figure 10, we noticed that the features in SAE is much more sparse than those in NCL.

Figure 11: The changes of active dimensions over **NCL** training.

A.2 Implementation of MDS

Based on the trained CLIP, CLIP+SAE, CLIP+NCL, and DeCLIP, we feed the test split of `cc3m-wds` dataset to these pretrained models, around 15k image-text pairs to calculate MDS, according to Eq.(1). The features are the last-layer output from the text and image encoder. We tried to calculate the normalization of z_i and

[6] https://github.com/openai/CLIP

[7] https://huggingface.co/datasets/pixparse/cc3m-wds

z_t, but found that it makes little difference to the final results. It could be attributed to the existing normalization technique in image and text encoders in CLIP.

A.3 Interpretability Evaluation

A.3.1 Implementation

Embedding models h for activated image/text samples. Our interpretability metrics, i.e., ***EmbSim*** and ***WinRate*** are based on the embeddings of active image/text samples by each feature. We need the embedding models to obtain these embeddings, i.e., Z^+ and Z^-. We use the Vision Transformer (ViT-B-16-224-in21k) for image embeddings and the Sentence Transformer (all-MiniLM-L6-v2) for text embeddings. The goal here is to derive the general and effective image and text embeddings, so we can also use the image encoder and text encoder from CLIP.

A.3.2 Results

A.4 Qualitative evaluation results

`CrossD` (the majority features) capture shared semantics across modalities. Different from modality-specific features, `TextD` and `ImgD`, `CrossD` features capture common concepts that could be expressed in both visual and language modalities. We randomly select two `CrossD` features and show their top activated images and texts. As shown in Figure 12, Feature-6 mostly activates scenes involving individuals performing activities, especially outdoor activities, and feature-47 captures general outdoor environments. The coherence across both modalities reflects successful alignment, which is consistent with multimodal training objectives.

MDS for different methods MDS with monosemanticity enhancements. With the monosemanticity-improving models (SAE and NCL), we hypothesize that modality purity will become more pronounced, making dominant modality assignments more meaningful. To validate this, we calculate the MDS and visualize the distributions of the three feature groups across models in Figure 13. Interestingly, we find that CLIP, which is only trained on an image-text contrastive learning objective, contains a spectrum of features with different modality dominance. Specifically, its distribution skews towards the image modality, and this trend is consistent across all models. DeCLIP, on the other hand, shows a more balanced and less centered distribution. This suggests that DeCLIP, through self-supervision, extracts more modality-specific features, which might be overlooked by pure vision-language contrastive models like CLIP. The extracted features on top of NCL and SAE also exhibit less skewness, with SAE showing the most balanced distribution, indicating its strong capability to extract diverse monosemantic features.

EmbSimi and WinRate for Monosemanticity measurement. Firstly, we show the complete results for ***EmbSmi*** and ***WinRate*** in the Table 4.

Models	***EmbSim***		***WinRate***	
Activated→	Image	Text	Image	Text
CLIP	0.11	0.45	0.65	0.59
DeCLIP	0.06	-0.07	0.61	0.46
CLIP+NCL	0.14	0.45	**0.71**	0.60
CLIP+SAE	**0.17**	**0.74**	0.60	**0.61**

Table 4: Average interpretability scores (by examining the top activated images/texts) for features extracted from VLMs.

The results of monosemanticity changes as training goes on. We show the results of monosemanticity score changes as training goes on for both NCL and SAE in Figure 14.

Models	**CLIP**	**DeCLIP**	**CLIP+NCL**	**CLIP+SAE**
		Mono is ***EmbSim***		
Visual Mono	-0.007	0.009	**0.043**	0.005
Textual Mono	-0.017	-0.001	**0.210**	0.146
		Mono is ***WinRate***		
Visual Mono	-0.007	0.005	0.002	**0.030**
Textual Mono	-0.069	-0.059	**0.018**	0.016

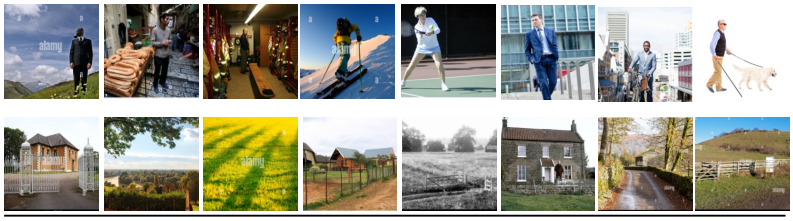
Table 5: The visual and textual monosemanticity. A higher value indicates that `ImgD` captures more visual than linguistic features, and vice versa for `TextD`.

A.5 Implementations and More Results for Case Studies

We provide the implementation details and more experimental results for the three case studies in the following.

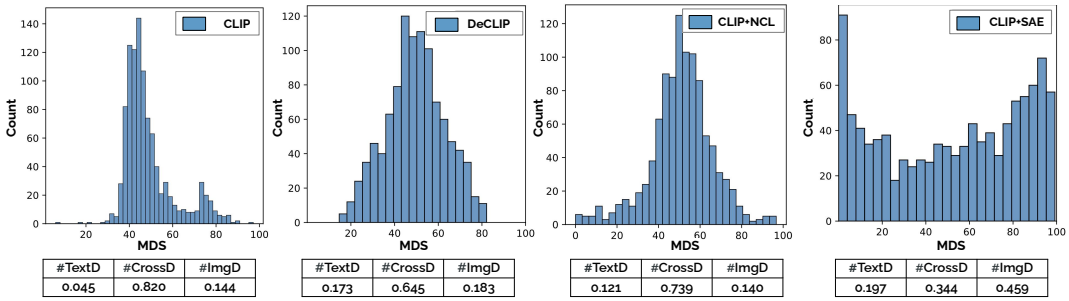
A.5.1 Case study 1: Understanding Gender Pattern in Different Modalities

Datasets. We select male and female images using a gender classifier touchtech/fashion-images-gender-age-vit-large-patch16-224-in21k-v3 from cc3m-wds validation set. We have both input images and text; the original gender classification accuracy is 83.4% and 73.4%, respectively.



Feature-6: Actions performed by individuals	*Feature-47: Outdoors Scenery*
Young man working on invention in a warehouse.	A stile on a public footpath overlooking the village on a frosty autumn morning.
cricketers exercise during a practice session.	A private chapel, and the wrought iron gates in the grounds.
Cricket player checks his bat during a training session.	Train track: a man blending in with the scenery as he stands on a railway track near a river
Basketball coach watches an offensive possession from the sideline during the second half.	surveying the scene: people look out over loch today on a warm day in the village

Figure 12: Activated images and texts by `CrossD` features. Top image row (feature 6): activities performed by individuals. Bottom image row (feature 47): scenery outside the doors. Text in blue aligns with visual concepts.



	#TextD	#CrossD	#ImgD
CLIP	0.045	0.820	0.144
DeCLIP	0.173	0.645	0.183
CLIP+NCL	0.121	0.739	0.140
CLIP+SAE	0.197	0.344	0.459

Figure 13: Modality Dominance Score (MDS) distributions of three feature categories for different VLMs.

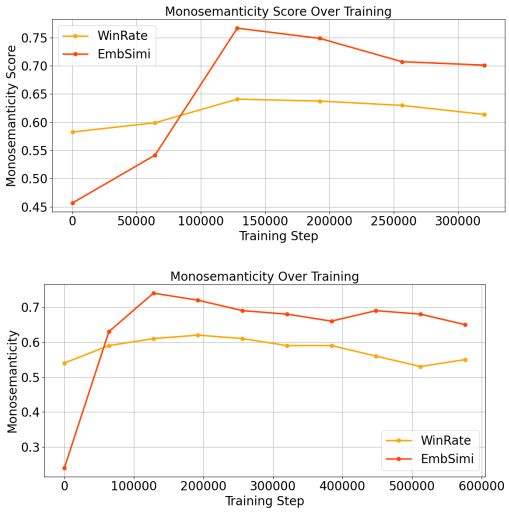


Figure 14: Monosemanticity (EmbSimi and WinRate) changes as training goes on. Upper is for CLIP+NCL, bottom is for CLIP+SAE.

Classification. As the intervened features are not compatible with existing pretrained text or text classifiers, we compare these features with the golden feature from male and female data. Specifically, we randomly select a female/male image with classification logits larger than 0.9 (ensuring the gender patterns are obvious) as the reference features. We use the same embedding models in §A.3, i.e., Vision Transformer and Sentence Transformer as the encoder and encode both intervened feature and golden feature. The intervened feature is labeled with the same label as the reference image, for which its distance in encoder space is smaller.

Intervention. There are different number of `ImgD` and `TextD` for a given representation of input sample. To avoid the effects of different numbers of removal features, we remove (set the corresponding dimension as zero) the minimal number between `ImgD` and `TextD`, and remove the same number of randomly selected features as a baseline.

`TextD` in male concepts. We also cluster different male descriptions according to the percentage of `TextD` features among all their top-20 activated features, and we calculate the frequency of the

top7 tokens in each cluster shown in Table 6. We remove the gendered personal pronouns, e.g., he, she, woman, man, boy, girl, and only focus on how gender-neutral concepts represent the gender. With more TextD injection, the textual descriptions become more sports-related, such as coach, basketball, soccer; while the sentences with less activated TextD have top words, such as party, hip, game, smile, home. This trend is consistent with the social stereotype that males are more active in sports activities.

A.5.2 Case Study 2: Defending Modality-Specific Adversarial Attacks

Models We employed the same ViT-B-32 CLIP as in §A.1 as the multimodality feature extractor shown in the Figure 7 to extract 1024-dimension features, so we use the categorized TextD, ImgD and CrossD calculated before. We use LLaVA-1.5-7b as the attacked VLM (Liu et al., 2023). The whole process of defensing adversarial attacks is two steps:

- **Generating adversarial images by injecting harmful requests.** We have a benign scenery image and a list of 50 harmful requests. Firstly, we create an image with a white background with the text saying the one piece of harmful request, as the contrast image. Then, we apply the alignment training by minimizing the distance between the benign image and the contrast image in the embedding space of the image encoder. The benign image is thus being injected with harmful semantics, denoted as $\mathbf{F}_{adv}$.
- **Defending the adversarial attacks.** To remove the toxicity of the adversarial samples, we employ the alignment training shown in Figure 7 by updating the embeddings of the adversarial samples. Specifically, we only select the target features, i.e., the ImgD, TextD, and CrossD to be involved in the training.

When attacking the VLM, we feed the adversarial images/samples along with the text prompt, i.e., the corresponding harmful request injected into the adversarial sample. For each adversarial sample, we repeat the attack process 100 times. For comparison, we apply the original generated 50 adversarial samples to attack VLM, and the average success rate is 73.26%; and the success rate of the (benign image - harmful request) is 10.00%. We conducted five independent runs for each experiment to ensure statistical reliability. Results in the tables show mean values across runs, with relative standard deviations below 3% for accuracy metrics.

Computing resources cost The experiments were conducted with a GPU with 48GB of memory. Adversarial sample generation requires approximately 4 GPU hours, while adversarial sample detoxification takes approximately 6 GPU hours.

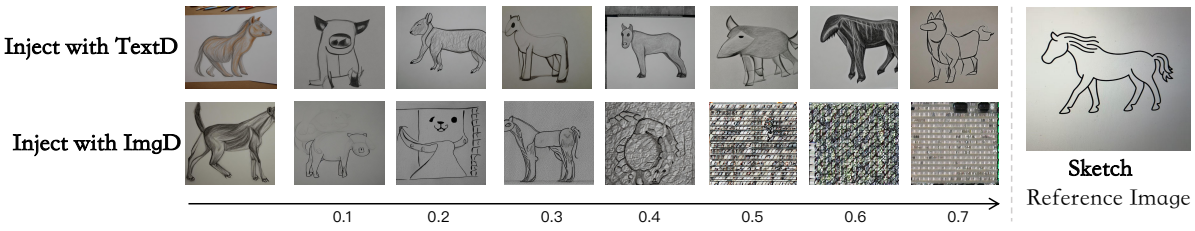
A.5.3 Case Study 3: Modality-Aware Control for Text-to-Image Generation

Models. We select Stable-Diffusion-v2 (https://huggingface.co/stabilityai/stable-diffusion-2) as our text-2-image generation model. As its image encoder (CLIP-ViT-H-14-laion2B-s32B-b79K) is not the same CLIP we used before, we recalculate the MDS distribution to derive the three categories of features.

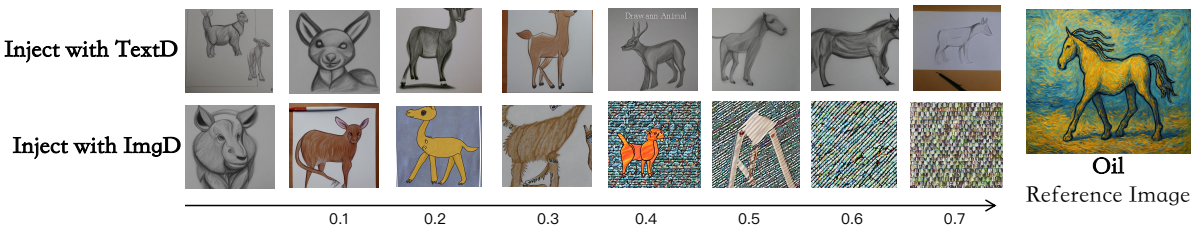
More results. We present additional images generated by modifying the original multimodal representation through feature injection from a reference image. To emphasize the distinction between ImgD and TextD, we use two reference images of horses in different backgrounds and artistic styles. Specifically, we compare two sets of images where features from sketch and oil painting styles are injected using ImgD. We observe that images influenced by sketches tend to be predominantly black and white, while those influenced by oil paintings appear more colorful. In contrast, the images generated using TextD remain visually similar across both the sketch and oil painting settings.

Percentage of TextD	Top8 words in male-related textual description
0.1	attends, party, hip, game, comedian, city, black, artist
0.12	smile, made, blue, outside, looks, home, got, book
0.18	artist, player, film, pop, performs, festival, young, suit
0.24	player, football, basketball, team, game, portrait, holding, gym

Table 6: Representative words in male-related descriptions with different percentages of TextD.



(1) Using the sketch reference image for a text-to-image experiment



(2) Using the oil reference image for a text-to-image experiment

Figure 15: Generated new images from the VLM with the text prompt *"Please draw an animal"* and varying levels of intervention from different reference images. We found that *TextD* captures significant semantic information, such as shape, etc. Notably, when a sketch is selected as the reference image, both *imgD* and *TextD* display sketch-like stylistic features. When oil-painting is chosen as the reference image, both *imgD* and *TextD* exhibit styles that resemble oil paintings. Comparatively, the stylistic differences between *imgD* in conditions (1) and (2) are distinct: *imgD* in (1) lacks color, whereas *imgD* in (2) presents diverse coloration. Similar to Figure 8, *TextD* does not affect low-level visual features, while *ImgD* shows significant distortion at higher α values.