

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

Anonymous ACL submission

Abstract

The success of vision-language models is primarily attributed to effective cross-modal alignment between vision and language. However, modality gaps persist even in well-aligned models and may be necessary for human perception, as evidenced by modality-specific phenomena such as visual texture and linguistic tone. These observations motivate us to computationally measure and leverage modality gaps to explore their utility in downstream applications. In this paper, we introduce the **Modality Dominance Score (MDS)**, which attributes multimodal features to specific modalities by categorizing them as vision-dominant, language-dominant, or cross-modal. We then propose automatic interpretability metrics to evaluate these modality-specific features in a scalable manner. Finally, we demonstrate how the identified modality-specific features enable training-free probing and editing methods for understanding model perception across genders, generating adversarial examples, and controlling text-to-image generation. Combined with task-agnostic interpretability tools, our work provides a systematic framework for analyzing and efficiently controlling multimodal models.

1 Introduction

Multimodal Models (MMs) have become foundational to the advancement of AI, enabling 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), a prominent class of MMs, 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 research aims to deepen our understanding of how different modalities interact and diverge within MMs (Liang et al., 2022; Rawal

et al., 2023; Schrodi et al., 2025; Zhang et al., 2025). For instance, Liang et al. (2022) revealed the modality gap as a geometric phenomenon in which image and text embeddings reside in disjoint regions of the shared embedding space. Zhang et al. (2025) proposed the MC² benchmark and identified clear modality bias across 18 tested large VLMs. Researchers have also developed automatic measurement methods (Liang et al., 2022; Rawal et al., 2023; Parcalabescu and Frank, 2023, 2025) and optimization techniques to calibrate modality gaps (Rawal et al., 2023; Zhang et al., 2024).

Despite these measurement and optimization methods, existing studies treat modality gaps as undesirable imperfections—primarily diagnosing model collapse (Rawal et al., 2023; Zhang et al., 2024; Schrodi et al., 2025; Parcalabescu and Frank, 2025) or motivating new training algorithms for improved alignment. Our work takes a different perspective: we posit that *modality gaps are both prevalent and beneficial for downstream tasks*. This perspective is grounded in cognitive science, where modality commonality and separation have long been central themes. Paivio (1991); Spence (2011); Fan et al. (2016) have examined how humans integrate and differentiate information across sensory modalities, 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 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 setting.

085 These metrics are scalable and compatible with
086 existing top-k activated interpretations.

087 3. We design lightweight probing and steering
088 methods to analyze models’ perception and
089 demonstrate how modality-specific features enable
090 control over VLM behavior in tasks such as
091 bias analysis, adversarial example generation, and
092 text-to-image generation.

093 2 Related Work

094 **Modality Gap.** Researchers have identified that
095 modality bias and gaps are prevalent in both early
096 VLMs (Liang et al., 2022) and large VLMs (Zhang
097 et al., 2025). Moreover, modality gaps have been
098 shown to negatively impact downstream tasks such
099 as video understanding (Rawal et al., 2023) and
100 object detection (Schrodi et al., 2025). Several
101 metrics have been proposed to quantify modality
102 gaps, including geometry-based methods such as
103 L2M (Liang et al., 2022) and RGM (Schrodi
104 et al., 2025), as well as Shapley-value-based ap-
105 proaches (Parcalabescu and Frank, 2023, 2025),
106 which measure the degree to which individual
107 modalities contribute to model predictions. *Our*
108 *work differs in two key aspects:* (i) we measure
109 modality gaps at the feature level, identifying how
110 individual features respond differentially to differ-
111 ent modalities; (ii) we leverage modality-specific
112 features for probing and steering, treating modality
113 specialization as a functional asset rather than an
114 imperfection.

115 **Interpretability Measurements.** Existing inter-
116 pretability measurements are based on summariz-
117 ing patterns in top-k activated samples. For exam-
118 ple, logit lens (nostalgebraist, 2020) has inspired
119 many studies in both unimodal and multimodal
120 representation understanding (Parekh et al., 2024;
121 Jiang et al., 2025). The embedding-based exten-
122 sion (Phukan et al., 2025) alleviates its limitation
123 in processing contextual-related concepts. More
124 recently, LLMs have been used to generate explana-
125 tions for activation patterns, with prediction accu-
126 racy on held-out samples serving as an interpretabil-
127 ity proxy (Bills et al., 2023). However, this LLM-
128 as-a-judge approach is computationally expensive
129 and only measures semantic coherence within the
130 unimodality, neglecting cross-modality consistency.
131 *Our interpretability metrics address these limita-*
132 *tions through scalable embedding-based computa-*
133 *tion and introduce modality fidelity to fill the gap.*

134 3 Identify Modality-Specific Features

135 Cross-modal alignment is important for VLMs,
136 yet modality gaps remain a widely observed phe-
137 nomenon (Liang et al., 2022; Schrodi et al., 2025).
138 Liang et al. (2022); Schrodi et al. (2025) attributed
139 this to geometric characteristics of the represen-
140 tation space; notably, Schrodi et al. (2025) ob-
141 served that modality bias can be traced to a few
142 embedding dimensions. Focusing on the represen-
143 tation space as well, we go further by attributing
144 each dimension to a specific modality dominance
145 class—text-dominant, image-dominant, or cross-
146 modal—enabling fine-grained analysis and control.

147 Background: Modality Alignment in VLMs.

148 Typically, there are an image encoder and a text
149 encoder in a VLM for image and text input pro-
150 cessing, respectively. Specifically, the image-text
151 pair $(x_{\text{img}}, x_{\text{txt}})$ is fed to an image encoder f_{img}
152 and a text encoder f_{txt} within the model, respectively
153 and the final-layer representations $z_{\text{img}} \in \mathbb{R}^D$ and
154 $z_{\text{txt}} \in \mathbb{R}^D$ are then optimized jointly in the shared
155 D -dimensional representation space. An alignment
156 loss, such as the contrastive loss in CLIP (Ilharco
157 et al., 2021) across the two modalities, is applied
158 for modality alignment.

159 3.1 Modality-specific Feature Identification

160 To measure the modality gap, Liang et al. (2022)
161 used the difference between the center of image
162 embeddings and text embeddings of M input pairs,
163 i.e., $\frac{1}{M} (\frac{1}{M} \sum_{i=1}^M \|z_{\text{img},i}\|^2 - \frac{1}{M} \sum_{i=1}^M \|z_{\text{txt},i}\|^2)$. We
164 extend this model-level measurement to a fine-
165 grained metric, i.e., the predominant modality as-
166 sociated with each dimension $d \in \{1, 2, \dots, D\}$
167 in the shared embedding space. The proposed
168 modality dominance score (MDS), denoted as $R(d)$
169 shown in Eq. (1) reflects how strongly the d -th fea-
170 ture¹ is influenced by the image modality:

$$171 R(d) = \frac{1}{M} \frac{\sum_{i=1}^M \|z_{\text{img},i}^{(d)}\|^2}{\sum_{i=1}^M \|z_{\text{img},i}^{(d)}\|^2 + \sum_{i=1}^M \|z_{\text{txt},i}^{(d)}\|^2}. \quad (1)$$

172 Specifically, we feed M image-text pairs to the
173 VLM and extract the corresponding image features
174 $z_{\text{img},i}$ and text features $z_{\text{txt},i}$ for i -th input. For each
175 d -th dimension in the D -dimension shared space,
176 we calculate the relative activation between the
177 features from the two modalities. This modality
178 fraction is averaged over more than $M = 10k$

¹Each feature dimension corresponds directly to a neu-
ron in the VLM’s final layer; our study thus focuses on the
interpretability of the model’s intrinsic components.

input pairs, providing a representative estimate of the modality distribution. Implementation details are in Appendix A.1.

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

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

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. This modality fidelity will be evaluated in §4.2.

3.2 Quantitative Evaluation for MDS

To verify that modality-specific features effectively capture their intended modality information, we employ an intervention-based evaluation. Specifically, we remove these features from the original CLIP ViT-H/14 (LAION-2B) (Ilharco et al., 2021) representation by zeroing out their corresponding indices, then use the modified representations as input to a logistic regression classifier. We evaluate performance on image and text classification tasks 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 features 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 features from the three groups, and then display their most-activated images and texts in Figure 1 (ImgD), Figure 2 (TextD), and Figure A1 (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

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

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

the modality alignment, while the visual commonalities 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 gave 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 § A.2.

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, which hinders interpretability. Monosemanticity (Elhage et al.; Bills et al., 2023; Gurnee et al.,

Feature-647: Pattern and others.	Feature-667: Scenes in winter and other.
A bed with tufted upholstery.	White trotting on snowy ground with a tree.
Seamless pattern, owers on a background.	Covering the trailhead in a winter wonderland.
Could new showroom and model signal the start of the	the image of drum under the white background.

Figure 1: Activated images and texts (in Table 1) by D. 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 1) by D. 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.

2023; Yan et al., 2024) has thus emerged as a Building upon the “soft” top-k activation frame- 286
paradigm for deriving interpretable features that enwork, we propose scalable, embedding-based eval- 287
code single, coherent concepts. However, scalingation metrics tailored to multimodal models. In 288
interpretability evaluation remains an open chalthis context, interpretability encompasses two di- 289
lence due to the heavy reliance on costly humammensions: 290
annotations (Gao et al., 2024) or LLM explana• Monosemanticity (within-modality coherence). 291
tions (Bills et al., 2023). To address this bottleneck, It measures whether a feature’s top-k activated 292
we propose a suite of automated metrics to measure samples exhibit semantic coherence within a single 293
feature interpretability. single modality based on their embedding similarity. 294

4.1 Overview

270
271 Top-k activation-based interpretation. A feature 272
is considered interpretable if its semantic mean- 273
ing can be readily understood by humans. In 274
practice, interpretability is assessed by examining 275
whether the top-k most highly activated samples 276
exhibit coherent and consistent patterns—a stan- 277
dard approach for analyzing both language (Geva, 278
et al., 2021) and multimodal models (Parekh et al., 279
2024). To incorporate semantic similarity, Bills 280
et al. (2023) extended this approach by computing 281
the correlation between predicted activation values 282
(based on explanations of the top-k samples) and 283
true activations. This correlation-based method re- 284
laxes the hard constraint of requiring perfect pattern 285
consistency across top-k samples.

286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301

302
303
304
305

306
307
308
309
310

$Z \in \mathbb{R}^{m \times d}$. 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 $EmbSim(z^{(d)})$ (EmbSim). We also propose a binary metric to avoid the different scales in different modalities, denoted as $W(z^{(d)})$ (WinRate):

$$EmbSim(z^{(d)}) = \frac{1}{m(m-1)} \sum_{i \neq j} \frac{(S_{+})_{ij} - (S_{-})_{ij}}{(S_{+})_{ij}};$$

$$W(z^{(d)}) = \frac{1}{m(m-1)} \sum_{i \neq j} \mathbb{1}_{[(S_{+})_{ij} > (S_{-})_{ij}]}.$$

The overall interpretability score is the average across all 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.²

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:

$$\text{Visual Fidelity} = \text{Mono}^{\text{vis}}(ImgD) - \text{Mono}^{\text{vis}}(TextD)$$

$$\text{Textual Fidelity} = \text{Mono}^{\text{txt}}(TextD) - \text{Mono}^{\text{txt}}(ImgD)$$

4.3 Interpretability Evaluation

Sparse Autoencoders (SAEs) (Cunningham et al., 2023) have been shown to effectively produce monosemantic features by enforcing sparsity constraints. To validate our evaluation framework, we apply our metrics to compare CLIP with and without SAE, expecting results consistent with existing literature on SAE’s interpretability benefits. Beyond this validation, we also investigate whether other representation learning methods can enhance feature interpretability.

4.3.1 Comparison Models

We incorporate several representation learning algorithms that aim to learn modality-specific features. Implementation details are provided in § A.3.

²We calculate the average of EmbSim and WinRate as Mono in the main content; the separate results for the two metrics can be found in § A.4.2.

Multimodal SAEs. SAEs have emerged as a scalable tool for transforming polysemantic features into interpretable, monosemantic ones across different LLMs (Templeton, 2024; Gao et al., 2024; Lieberum et al., 2024). We extend it for VLMs by training a single SAE model $Z \rightarrow Z$ 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 TopK 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} :

$$z^{sae} = \text{TopK}(W_{enc} z, b_{pre});$$

$$\hat{z} = W_{dec} z^{sae} + b_{pre}.$$

$z \in \mathbb{R}^D$ is the inputs of SAE, i.e. z_{img} or z_{txt} . $z^{sae} \in \mathbb{R}^n$ is the learned sparse representation. We train the multimodal SAE to reconstruct z_t . DeCLIP. Beyond multimodal supervision (image-text pairs), DeCLIP (Li et al., 2022) also incorporates single-modality 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,

$$E_{z_{img}, z_{txt}} \log \frac{\exp(g(z_{img})) > g(z_{txt})}{E_{z_{txt}} \exp(g(z_{img})) > g(z_{txt})};$$

where here we use a ReLU-activated MLP network to map input features to non-negative outputs.

4.3.2 Evaluation Results

Results of Monosemanticity. We compute the Mono 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 (both enforce the feature sparsity) exhibit the highest overall monosemanticity for both activated input images and texts. (ii) DeCLIP does not enhance interpretability through self-supervision alone; the monosemanticity on the textual side becomes even worse. This suggests that polysemantic features remain prevalent in DeCLIP.

Moreover, we observe that monosemanticity enhancement encourages more modality-specific

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

neurons in Figure 3. The figure shows the MDS distributions across CLIP and other variants. CLIP contains a spectrum of features with different modality dominance, with its distribution skewed 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.

Figure 4: Monosemanticity.

Figure 5: Modality Fidelity.

Leveraging the Modality Gap

Beyond interpretability, we design lightweight probing and steering methods based on modality-specific features to analyze VLMs' perceptual preferences and enable precise behavioral control. Implementation details are in § A.5.

5.1 Understanding Gender Patterns

Gender is represented using both visual and textual features, and these data are used to train VLMs. To examine whether gender exhibits modality-specific representations, we investigate whether feminine concepts are more frequently conveyed through visual cues (e.g, via more colorful clothing) than through textual descriptions.

To answer this question, 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.

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.

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 (as the removal of ImgD from the image leads to larger classification degradation), whereas male concepts are more affected by the removal of TextD.

We sample female images that differ in the proportion of their most activated features categorized as `ImgD` features. The results are shown in Figure 6. From left to right, the fraction of activated `ImgD` features increases, and the images exhibit progressively more (stereotypically) feminine visual details, such as a backless skirt and hair accessories. The middle images show professional women, such as a doctor; while the leftmost image shows only a pair of legs in sports shoes, with minimal feminine cues aside from the pink color.

Figure 6: Female images ordered by increasing percentages of `ImgD` (0.14, 0.16, 0.18, 0.20, 0.22, 0.24), corresponding to more feminine visual concepts.

5.2 Defense against Adversarial 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).

An adversarial sample is a benign-appearing image (e.g., a landscape) injected with harmful semantic content, such as the phrase “I want to make a bomb”. One defense-oriented optimization strategy involves minimizing the distance between the embeddings of an adversarial sample F_{adv} , and a paired benign sample F_{ben} , by updating the features in the adversarial sample (in Figure 7). The paired benign image is injected with a friendly phrase, such as “peace and love”.

To isolate the effects of our identified modality-specific features, we restrict alignment training to a selected set of feature indices corresponding to `ImgD`, `TextD`, and `CrossD`. The alignment loss is defined as $L = k \| F_{adv}[:, I] - F_{ben}[:, I] \|_2$. The optimized adversarial sample is then used to attack a VLM, LLaVA-1.5-7b (Liu et al., 2023). We evaluate the VLM’s responses using an LLM-as-a-Judge framework, where DeepSeek-V3 (DeepSeek AI et al., 2024) produces a binary label indicating attack success. We hypothesize that feature sets encoding richer malicious semantics contribute most strongly to effective adversarial defense. Results. The attack success rates are shown in Table 3. For alignment training, we select an equal number of features from `ImgD`, `TextD`, and `CrossD` and additionally include a baseline that randomly

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.

samples the same number of features from the union of all feature sets. Each adversarial sample is used to attack the VLM 100 times, and we generate a total of 50 adversarial samples. We observe that (i) compared with the original adversarial samples, alignment training using any subset of selected features reduces the attack success rate, indicating that feature-level alignment provides a degree of defense; (ii) using `TextD` yields the best defense performance, followed by `CrossD` and `ImgD`. This ordering is consistent with the nature of the attack, as the injected adversarial content primarily arises from harmful textual semantics. These results demonstrate that `TextD` features effectively capture most of the semantic content relevant to the attack. In contrast, `CrossD` captures only partial semantic information, while `ImgD` is least correlated with semantic information, resulting in minimal benefits for such modality-specific jail-break defense.

Target feature	<code>ImgD</code>	<code>TextD</code>	<code>CrossD</code>
Success Rate (#)	62.71%	24.89%	35.44%

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

5.3 Controllable Text-to-Image Generation

Despite the impressive capabilities of text-to-image generation models (Koh et al., 2024; Swamy et al., 2024), their internal mechanisms for translating linguistic semantics into visual details remain poorly understood. A key challenge lies in disentangling the roles of modality-specific features in shaping generation fidelity and controllability. To address this, we conduct a feature intervention experiment during the image generation process of Stable Diffusion v2 (Rombach et al., 2022). Intervention. As shown in Figure 9, we investigate the generation process by intervening in modality-

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 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.

specific features in Stable-Diffusion-v2, which can be viewed as a VLM with an encoder and decoder (generator). The input text prompt is “Please draw an animal”. The encoder generates an embedding T , representing the original multimodal embedding T^0 to the generator of the VLM with different interpolation weights 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 of 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 by TextD typically depict clear main subjects without transferring visual background details from the reference image. In contrast, interventions resulting multimodal embedding is computed as $T^Q[I] = T[I] + (1 - \alpha)R[I]$, where the interpolation is only applied to the index set corresponding to TextD, CrossD, and ImgD features.

Figure 9: Controllable text-2-image generation via editing the modality-specific information from the reference image.

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.

583	Limitation	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.	636 637 638 639 640 641 642								
584	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.	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.	643 644 645 646								
585		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.	647 648 649 650 651 652								
586			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.	653 654 655 656 657							
587				Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. arXiv preprint arXiv:2102.03334.	658 659 660 661						
588					Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. 2024. Generating images with multimodal language models. Advances in Neural Information Processing Systems, 36.	662 663 664 665 666					
589						Yanguang 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.	667 668 669 670 671				
590							Zhaowei Li, Wei Wang, YiQing Cai, Qi Xu, Pengyu Wang, Dong Zhang, Hang Song, Botian Jiang, Zhida Huang, and Tao Wang. 2025. Uni edmlm: 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.	672 673 674 675 676 677 678			
591								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.	679 680 681 682		
592									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.	683 684 685 686 687	
593										Tom Lieberum, Senthoran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse	688 689 690 691
594											
595											
596											
597											
598											
599											
600	References										
601	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.										
602		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.									
603			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.								
604				DeepSeek-AI, Aixin Liu, and Bei Feng et al. 2024. Deepseek-v3 technical report . Preprint, arXiv:2412.19437.							
605					Nelson Elhage, Neel Nanda, Catherine Olsson, and Others. A mathematical framework for transformer circuits . Transformer Circuits Thread (2022).						
606						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 connectonal architecture . Cerebral Cortex, 26:bhw157.					
607							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.				
608											
609											
610											
611											
612											
613											
614											
615											
616											
617											
618											
619											
620											
621											
622											
623											
624											
625											
626											
627											
628											
629											
630											
631											
632											
633											
634											
635											

692	autoencoders everywhere all at once on gemma 2.	Albuquerque, New Mexico. Association for Computational Linguistics.	747
693			748
694	Tsung-Yi Lin, Michael Maire, Serge Belongie, James	Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya	749
695	Hays, Pietro Perona, Deva Ramanan, Piotr Dollár,	Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sas-	750
696	and C Lawrence Zitnick. 2014. Microsoft coco:	try, Amanda Askell, Pamela Mishkin, Jack Clark, and	751
697	Common objects in context. In European confer-	1 others. 2021. Learning transferable visual models	752
698	ence on computer vision, pages 740–755. Springer.	from natural language supervision. arXiv preprint	753
		arXiv:2103.00020.	754
699	Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae		
700	Lee. 2023. Improved baselines with visual instruc-	Isha Rawal, Shantanu Jaiswal, Basura Fernando, and	755
701	tion tuning.	Cheston Tan. 2023. Dissecting multimodality in videoqa transformer models by impairing modality fusion . In International Conference on Machine Learning.	756
702	Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee.		757
703	2019. Vilbert: Pretraining task-agnostic visiolinguistic		758
704	representations for vision-and-language tasks. In		759
705	Advances in Neural Information Processing Systems	Robin Rombach, Andreas Blattmann, Dominik Lorenz,	760
706	pages 13–23.	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.	761
707	Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan		762
708	Nam, Honglak Lee, and Andrew Y Ng. 2011. Multi-		763
709	modal deep learning. In Proceedings of the 28th In-		764
710	ternational Conference on Machine Learning (ICML-		765
711	11), pages 689–696.	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.	766
712	nostalgebraist. 2020. Interpreting gpt		767
713	the logit lens. https://www.less-		768
714	wrong.com/posts/AcKRB8wDpdaN6v6ru/		769
715	interpreting-gpt-the-logit-lens .		770
716	Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel	Piyush Sharma, Nan Ding, Sebastian Goodman, and	772
717	Goh, Michael Petrov, and Shan Carter. 2020. Zoom	Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of ACL.	773
718	in: An introduction to circuits. <i>Distill</i> , 5(3):e00024–		774
719	001.		775
720	Allan Paivio. 1991. Dual coding theory: Retrospect and	Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh.	776
721	current status. <i>Canadian Journal of Psychology/Re-</i>	2024. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models . In The Twelfth International Conference on Learning Representations.	777
722	<i>vue canadienne de psychologie</i> , 45(3):255.		778
723	Letitia Parcalabescu and Anette Frank. 2023. Mm-shap:		779
724	A performance-agnostic metric for measuring multi-		780
725	modal contributions in vision and language models	Charles Spence. 2011. Crossmodal correspondences: A tutorial review. <i>Attention, Perception, & Psychophysics</i> , 73(4):971–995.	781
726	& tasks. In Proceedings of the 61st Annual Meet-		782
727	ing of the Association for Computational Linguistics		783
728	(Volume 1: Long Papers), pages 4032–4059.	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.	784
729	Letitia Parcalabescu and Anette Frank. 2025. Do vision		785
730	& language decoders use images and text equally?		786
731	how self-consistent are their explanations? In The		787
732	Thirteenth International Conference on Learning		788
733	Representations.		789
734	Jayneel Parekh, Pegah Khayatan, Mustafa Shukor, Alas-	Vinitra Swamy, Malika Satayeva, Jibril Frej, Thierry Bossy, Thijs Vogels, Martin Jaggi, Tanja Käser, and Mary-Anne Hartley. 2024. Multimodal—multimodal, multi-task, interpretable modular networks. <i>Advances in Neural Information Processing Systems</i> , 36.	791
735	dair Newson, and Matthieu Cord. 2024. A concept-		792
736	based explainability framework for large multimodal		793
737	models. <i>Advances in Neural Information Processing</i>		794
738	<i>Systems</i> , 37:135783–135818.		795
739	Anirudh Phukan, Divyansh, Harshit Kumar Morj, Vaish-		796
740	navi, Apoorv Saxena, and Koustava Goswami. 2025.		
741	Beyond logit lens: Contextual embeddings for robust hallucination detection & grounding in VLMs .	Adly Templeton. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. Anthropic.	797
742	In Proceedings of the 2025 Conference of the Na-		798
743	tions of the Americas Chapter of the Association for		799
744	Computational Linguistics: Human Language Tech-		
745	nologies (Volume 1: Long Papers), pages 9661–9675,	Yifei Wang, Qi Zhang, Yaoyu Guo, and Yisen Wang.	800
746		2024. Non-negative contrastive learning. ICLR.	801

802 Hanqi Yan, Yanzheng Xiang, Guangyi Chen, Yifei
803 Wang, Lin Gui, and Yulan He. 2024. [Encourage](#)
804 [or inhibit monosemanticity? revisit monosemantic-](#)
805 [ity from a feature decorrelation perspective](#). ArXiv,
806 [abs/2406.17969](#).

807 Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Jin-
808 guo Zhu, Han Liu, Jinghui Chen, Ting Wang, and
809 Fenglong Ma. 2024. Vlattack: Multimodal adversar-
810 ial attacks on vision-language tasks via pre-trained
811 models. Advances in Neural Information Processing
812 Systems, 36.

813 Jinghao Zhang, Guofan Liu, Qiang Liu, Shu Wu, and
814 Liang Wang. 2024. Modality-balanced learning for
815 multimedia recommendation. In Proceedings of the
816 32nd ACM International Conference on Multimedia,
817 pages 7551–7560.

818 Yu Zhang, Jinlong Ma, Yongshuai Hou, Xuefeng Bai,
819 Kehai Chen, Yang Xiang, Jun Yu, and Min Zhang.
820 2025. Evaluating and steering modality preferences
821 in multimodal large language model. arXiv preprint
822 [arXiv:2505.20977](#).

823	A Appendix	(around 2900k image-text pairs) from cc3m-wds ⁴ .	871	
824	A.1 Implementation of MDS	We train the two variants, i.e., SAE and NCL, on top of the pretrained CLIP using a single 3090 GPU.	872	
825	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.	DeCLIP. We use the checkpoint released in https://github.com/Sense-GVT/DeCLIP to extract the last layer features, z_i and z_t .	873	
826			874	
827			875	
828			876	
829			877	
830				
831	We tried to calculate the normalization of z_i and 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.	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 $m = d$ 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.	878	
832			879	
833			880	
834			881	
835			882	
836	A.2 CrossD Qualitative Results		883	
837	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 A1, 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.	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	885	
838			886	
839			887	
840			888	
841			889	
842			890	
843			891	
844			892	
845		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 A2.		893
846			894	
847			895	
848			896	
849			897	
850			898	
851	A.3 Implementation for Comparison Models		899	
852	The three comparison models, DeCLIP, Multimodal SAE, and Multimodal NCL are all on top of the canonical ViT-B-32 CLIP ³ 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 <code>open_clip.create_model_and_transforms</code> function in the published https://github.com/mlfoundations/open_clip .		900	
853			901	
854			902	
855			903	
856			904	
857			905	
858			906	
859			907	
860			908	
861			909	
862			910	
863			911	
864			912	
865			913	
866			914	
867			915	
868			916	
869			917	
870				

³<https://github.com/openai/CLIP>

⁴<https://huggingface.co/datasets/pixparse/cc3m-wds>

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 A1: Activated images and texts by 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.

we can also use the image encoder and text encoder from CLIP.

A.4.2 Results

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 A4. 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.

Figure A2: The changes of active dimensions over SAE training.

Figure A3: The changes of active dimensions over NCL training.

A.4 Interpretability Evaluation

A.4.1 Implementation

Embedding modelsh 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

EmbSmi and WinRate for Monosemanticity measurement. Firstly, we show the complete results for EmbSmi and WinRate in the Table A1. The results of monosemanticity changes as training goes on. We show the results of monosemanticity score changes as training goes on for both

Figure A4: Modality Dominance Score (MDS) distributions of three feature categories for different VLMs.

Models	EmbSim		WinRate	
	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

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
Models	Mono is WinRate			
	Visual Mono	-0.007	0.005	0.002
Textual Mono	-0.069	-0.059	0.018	0.016

Table A1: Average interpretability scores (by examining the top activated images/texts) for features extracted from VLMs.

Table A2: The visual and textual monosemanticity. A higher value indicates that D captures more visual than linguistic features, and vice versa for TextD.

cc3m-wds validation set. We have both input images and text; the original gender classification accuracy is 83.4% and 73.4%, respectively.

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.4, i.e., Vision Transformer and Sentence Transformer as the encoder and encode both intervened feature and golden feature. The intervened feature is labeled

Figure A5: Monosemanticity (EmbSim and WinRate) with the same label as the reference image, for changes as training goes on. Upper is for CLIP+NCL which its distance in encoder space is smaller, bottom is for CLIP+SAE.

NCL and SAE in Figure A5.

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 Understanding Gender Patterns

Datasets. We select male and female images using a gender classifier [touchtech/fashion-images-gender-age-vit-large-patch16-224-in21k-v3](#) from

Intervention. There are different number of I_{imgD} and T_{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 I_{imgD} and T_{textD} , and remove the same number of randomly selected features as a baseline.

T_{textD} in male concepts. We also cluster different male descriptions according to the percentage of T_{textD} features among all their top-20 activated features, and we calculate the frequency of the top7 tokens in each cluster shown in Table A3. We remove the gendered personal pronouns, e.g., he,

she, woman, man, boy, girl, and only focus on Computing resources cost The experiments
 how gender-neutral concepts represent the gender were conducted with a GPU with 48GB of mem-
 With moreTextDinjection, the textual descriptions ory. Adversarial sample generation requires ap-
 become more sports-related, such as coach, basketroximately 4 GPU hours, while adversarial sample
 ball, soccer; while the sentences with less activatedetoxi cation takes approximately 6 GPU hours.
 TextD have top words, such as party, hip, game, A.5.3 Controllable Text-to-Image Generation
 smile, home. This trend is consistent with the so- Models. We select Stable-Diffusion-v2
 social stereotype that males are more active in sports Models. We select Stable-Diffusion-v2
 activities. ([https://huggingface.co/stabilityai/](https://huggingface.co/stabilityai/stable-diffusion-2)
[stable-diffusion-2](https://huggingface.co/stabilityai/stable-diffusion-2)) as our text-2-image
 generation model. As its image encoder (CLIP-

A.5.2 Defense against Adversarial Attacks

We employed the same ViT-B-32 CLIP ViT-H-14-laion2B-s32B-b79K) is not the same
 as in §A.3 as the multimodality feature extractorCLIP we used before, we recalculate the MDS
 shown in the Figure 7 to extract 1024-dimensiondistribution to derive the three categories of
 features, so we use the categorizedTextD, ImgD features.

andCrossDcalculated before. We use LLaVA-1.5- More results. We present additional images gen-
 7b as the attacked VLM (Liu et al., 2023). The erated by modifying the original multimodal rep-
 whole process of defending adversarial attacks isresentation through feature injection from a refer-
 two steps: ence image. To emphasize the distinction between
 ImgDandTextD, we use two reference images of

- Generating adversarial images by inject- ing harmful requests. We have a benign
 scenery image and a list of 50 harmful re- requests. Firstly, we create an image with a
 white background with the text saying the one piece of harmful request, as the contrast im-
 age. Then, we apply the alignment training by minimizing the distance between the benign
 image and the contrast image in the embed- ding space of the image encoder. The benign
 image is thus being injected with harmful se- mantics, denoted as F_{adv} .
- Defending the adversarial attacks. To re- move the toxicity of the adversarial samples,
 we employ the alignment training shown in Figure 7 by updating the embeddings of the
 adversarial samples. Speci cally, we only se- lect the target features, i.e., t_{regD} TextD,
 and CrossD to be involved in the training.

When attacking the VLM, we feed the adver- sarial images/samples along with the text prompt,
 i.e., the corresponding harmful request injected into the adversarial sample. For each adversarial sam-
 ple, we repeat the attack process 100 times. For comparison, we apply the original generated 50 ad-
 versarial 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 ve independent runs for each experi-
 ment to ensure statistical reliability. Results in the tables show mean values across runs, with relative
 standard deviations below 3% for accuracy metrics.

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 A3: Representative words in male-related descriptions with different percentages of TextD.

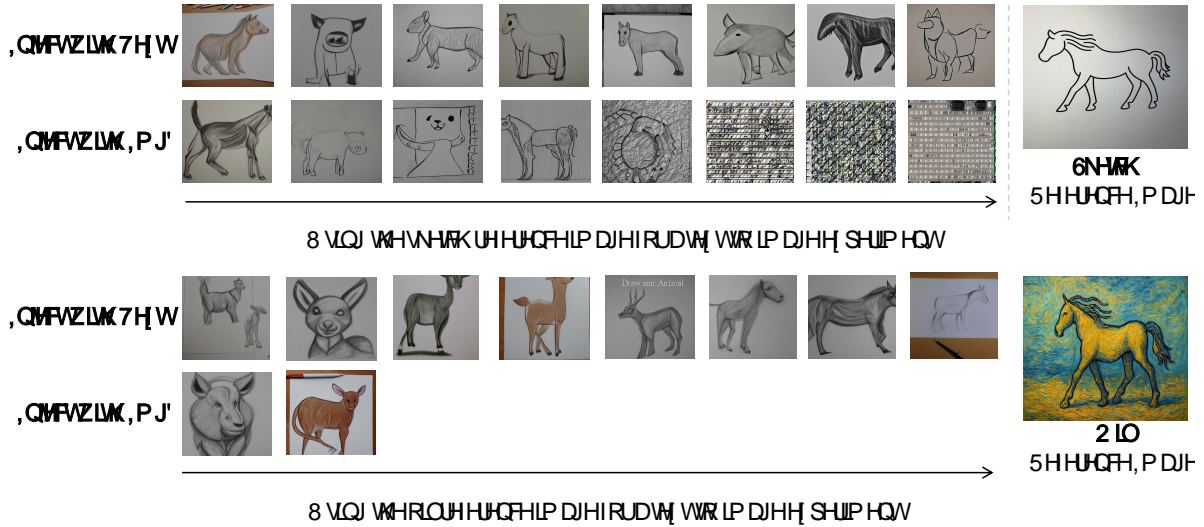


Figure A6: Generated new images from the VLM with the text prompt "Please draw an animal" and varying levels of intervention from different reference images. We found that *TextD* captures significant semantic information, such as shape, etc. Notably, when a sketch is selected as the reference image, 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 values.