A Survey on Mechanistic Interpretability for Multi-Modal Foundation Models

Anonymous ACL submission

Abstract

The rise of foundation models has transformed machine learning research, prompting efforts to uncover their inner workings and develop more efficient and reliable applications for better control. While significant progress has been made in interpreting Large Language Models (LLMs), multimodal foundation models (MMFMs)—such as contrastive vision-language models, generative vision-language models, and text-to-image models—pose unique interpretability challenges beyond unimodal frameworks. Despite initial studies, a substantial gap remains between the interpretability of LLMs and MMFMs. This survey explores two key aspects: (1) the adaptation of LLM interpretability methods to multimodal models and (2) understanding the mechanistic differences between unimodal language models and cross-modal systems. By systematically reviewing current MMFM analysis techniques, we propose a structured taxonomy of interpretability methods, compare insights across unimodal and multimodal architectures, and highlight critical research gaps.

1 Introduction

The rapid development and adoption of multimodal foundation models (MMFMs)—particularly those integrating image and text modalities—have enabled a wide range of real-world applications. For example, text-to-image models (Rombach et al., 2022; Ramesh et al., 2022; Podell et al., 2023) facilitate image generation and editing, generative vision-language models (VLMs) (Zhu et al., 2023; Agrawal et al., 2024) support tasks like visual question answering (VQA) or image captioning tasks, and contrastive (i.e., non-generative) VLMs such as CLIP (Radford et al., 2021) are widely used for image retrieval. As multimodal models advance, there is a growing need to understand their internal mechanisms and decision-making processes (Basu et al., 2024a). Mechanistic interpretability is crucial not only for explaining model behavior but also for enabling downstream applications such as model editing (Basu et al., 2024a), mitigating spurious correlations (Balasubramanian et al., 2024), and improving compositional generalization (Zarei et al., 2024).

Interpretability in machine learning, LLMs, and multimodal models is a broad and context-dependent concept, varying by task, objective, and stakeholder needs. In this survey, we adopt the definition proposed by Murdoch et al. (2019): "*The process of extracting and elucidating the relevant knowledge, mechanisms, features, and relationships a model has learned, whether encoded in its parameters or emerging from input patterns, to explain how and why it produces outputs.*" What constitutes "relevant knowledge" depends on the application. In memory editing, interpretability enables precise modifications to internal representations without disrupting other model functions. In attack detection, it highlights input features and activations signaling adversarial inputs. This survey examines interpretability methods through this lens, exploring how they uncover model mechanisms, facilitate practical applications, and reveal key research challenges.

While interpretability research has made significant progress in unimodal large language models (LLMs) (Meng et al., 2022a; Marks et al., 2024), the study of MMFMs remains comparatively underexplored. Given that most multimodal models are transformer-based, several key questions arise: *Can LLM interpretability methods be adapted to multimodal models*? If so, do they yield similar insights? *Do multimodal models exhibit fundamental mechanistic differences from unimodal language models*? Additionally, to analyze multimodal-specific processes like cross-modal interactions, *are entirely new methods required*? Finally, we also examine the practical impact of interpretability by asking—*How can multimodal interpretability methods enhance downstream applications*?

To address these questions, we conduct a comprehensive survey and introduce a three-dimensional taxonomy for mechanistic interpretability in multimodal models: (1) **Model Family** – covering text-to-image diffusion models, generative VLMs, and non-generative VLMs ; (2) **Interpretability Techniques** – distinguishing between methods adapted from unimodal LLM research and those originally designed for multimodal models; and (3) **Applications** – categorizing real-world tasks enhanced by mechanistic insights.[1] Our survey synthesizes existing research and uncovers the following insights: (i) LLM-based interpretability methods can be extended to MMFMs with moderate adjustments, particularly when treating visual and textual inputs similarly. (ii) Novel multimodal challenges arise such as interpreting visual embeddings in human-understandable terms, necessitating new dedicated analysis methods. (iii) While interpretability aids downstream tasks, applications like hallucination mitigation and model editing remain underdeveloped in multimodal models compared to language models. These findings can guide future research in multimodal mechanistic interpretability.

The **summary of our contributions** are:

- We offer a comprehensive survey of *mechanistic interpretability for multimodal foundation models* spanning generative VLMs, contrastive VLMs, and text-to-image diffusion models.
- We introduce a *simple and intuitive taxonomy* which helps to distinguish the mechanistic methods, findings, and applications across unimodal and multimodal foundation models, highlighting critical research gaps.
- Based on the mechanistic differences between LLMs and multimodal foundation models, we identify fundamental *open challenges and limitations* in multimodal interpretability, providing directions for future research

2 LLM Interpretability Methods for Multimodal Models

We first examine mechanistic interpretability methods originally developed for large language models and their adaptability to multimodal models with minimal to moderate modifications. Our focus is on ***how existing LLM interpretability techniques can provide valuable mechanistic insights into multimodal models.***

2.1 Linear Probing

Probing trains lightweight classifiers on *supervised* probing datasets, typically linear probes, on frozen LLM representations to assess whether they encode linguistic properties such as syntax, semantics, and factual knowledge (Hao et al., 2021; Liu et al., 2024e; Zhang et al., 2024b; Liu et al., 2023b; Beigi et al., 2024). This approach has been extended to multimodal models, introducing new challenges such as disentangling the relative contributions of each modality (i.e., visual or textual). To tackle these challenges, Salin et al. (2022) developed probing methods to specifically assess how Vision-Language models synthesize and merge visual inputs with textual data to enhance comprehension, while Dahlgren Lindström et al. (2020) investigated the processing of linguistic features within image-caption pairings in visual-semantic embeddings. Unlike in LLMs, where upper layers predominantly encode abstract semantics (Jawahar et al., 2019; Tenney et al., 2019), multimodal probing studies (Tao et al., 2024; Salin et al., 2022) suggest that intermediate layers in multimodal models are more effective at capturing global cross-modal interactions, whereas upper layers often emphasize local details or textual biases. Furthermore, despite the fact that probing applications in LLMs are centered on specific linguistic analyses, the scope of probing in multimodal models extends to more varied aspects. For instance, Dai et al. (2023) investigated object hallucination in vision-language models, analyzing how image encodings affect text generation accuracy and token alignment.

Main Findings and Gap. The main drawback of linear probing is the requirement of supervised probing data and training a separate classifier for understanding concept encoding in layers. Therefore, scaling it via multimodal probing data curation and training separate classifiers across diverse multimodal models is a challenge.

2.2 Logit Lens

The Logit Lens is an *supervised* interpretability method used to understand the inner workings of

[1]A detailed discussion of this taxonomy is provided in Sec. (B) in Appendix.

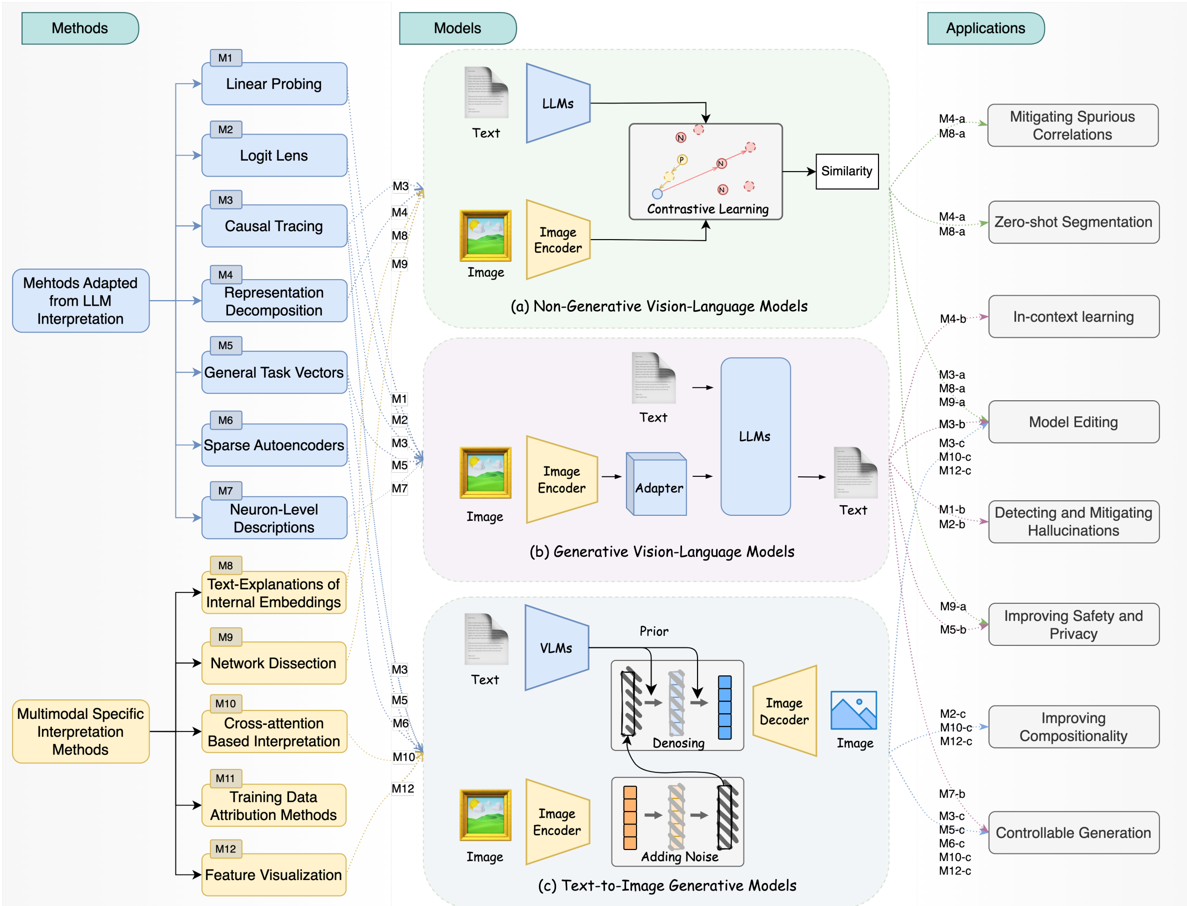


Figure 1: **In our survey, we study two types of mechanistic interpretability: (1) methods that adapted from LLM interpretability techniques and (2) multimodal-specific interpretability methods**. Different analysis methods are applied to three multimodal model architectures: (a) Non-generative Vision-Language Models, (b) Multimodal Large Language Models, and(c) Text-to-Image Generative Models (diffusion models especially). The interpretability insights from different methods and models can illuminate specific applications.

LLMs by examining the logits value of the output. This method conducts a layer-by-layer analysis, tracking logits at each layer (by projecting to the vocabulary space using the unembedding projection matrix) to observe how predictions evolve across the network. By decoding intermediate representations into a distribution over the output vocabulary, it reveals what the network "thinks" at each stage (nos, 2020; bel, 2023). In the context of multimodal modesl, studies show that predictions from earlier layers often exhibit greater robustness to misleading inputs compared to final layers (Halawi et al., 2024). Studies also demonstrate that anomalous inputs alter prediction trajectories, making this method a useful tool for anomaly detection (Halawi et al., 2024; bel, 2023). Additionally, for easy examples—situations where the model can confidently predict outcomes from initial layers—correct answers often emerge in early layers, enabling computational efficiency through adaptive early exiting (Schuster et al., 2022; Xin et al., 2020). Furthermore, the Logit Lens has been extended to analyze multiple inputs. Huo et al. (2024) adapted it to study neuron activations in feedforward network (FFN) layers, identifying neurons specialized for different domains to enhance model training. Further research has integrated contextual embeddings to improve hallucination detection (Phukan et al., 2024; Zhao et al., 2024a). Additionally, the "attention lens" introduced in (Jiang et al., 2024c) examines how visual information is processed, revealing that hallucinated tokens exhibit weaker attention patterns in critical layers.

Main Findings and Gap. Beyond multimodal language models, logit-lens can be potentially utilised to mechanistically understand modern models such as unified understanding and generation models such as (Xie et al., 2024a; Team, 2024).

2.3 Causal Tracing

Unlike passive diagnostic tools, Causal Tracing Analysis (Pearl, 2014) is rooted in causal inference that studies the change in a response variable following an active intervention on intermediate variables of interest (mediators). The approach has been widely applied to language models to pinpoint the network components—such as FFN layers—that are responsible for specific tasks (Meng et al., 2022a,b; Pearl, 2001). For instance, Meng et al. (2022a) demonstrated that mid-layer MLPs in LLMs are crucial for factual recall, while Stolfo et al. (2023) identified the important layers for mathematical reasoning. Building on this technique and using a *supervised* probing dataset, Basu et al. (2023) found that, unlike LLMs, visual concepts (e.g., style, copyrighted objects) are distributed across layers in the noise model for diffusion models, but can be localized within the conditioning text-encoder. Further, Basu et al. (2024b) identified critical cross-attention layers that encode concepts like artistic style and general facts. Recent works have also extended causal tracing to mechanistically understand generative VLMs for VQA tasks (Basu et al., 2024a; Palit et al., 2023; Yu and Ananiadou, 2024c), revealing key layers that guide model decisions in VQA tasks.

Main Findings and Gap. While causal tracing has been extensively used to analyze factuality and reasoning in LLMs, its application in multimodal models remains relatively limited. Expanding this method to newer, more complex multimodal architectures and diverse tasks remains an important challenge to address.

2.4 Representation Decomposition

A key property of transformer models is that layer-wise representations can be decomposed into a sum of preceding layers, enabled by the residual stream. This property is leveraged to extract circuit graphs in LLMs (Syed et al., 2023; Wang et al., 2022b; Conmy et al., 2023b; Basu et al., 2025). Circuit nodes, such as attention heads and MLP layers, can be further analyzed for the properties they encode (e.g., an attention head can encode *color* information). In multimodal models, representation decomposition has been instrumental in analyzing modality processing and layer-specific properties. Studies such as (Gandelsman et al., 2024a; Balasubramanian et al., 2024) leverage *supervised* probing datasets and propose a hierarchical decomposition approach—spanning layers, attention heads, and tokens—to provide granular insights into model behavior.

Layer-wise decomposition reveals that shallow layers primarily integrate modality-specific inputs into a unified representation, while deeper layers refine task-specific details through denoising (Yin et al., 2024). Tao et al. (2024) further demonstrated that intermediate layers capture broader semantic information, balancing modality-specific details with holistic understanding—crucial for tasks such as visual-language entailment. In diffusion models like Stable Diffusion, Prasad et al. (2023) found that lower U-Net layers drive semantic shifts, while higher layers focus on denoising, progressively refining the latent representations into high-quality outputs. Quantmeyer et al. (2024) utilized causal tracing with representation decomposition to identify CLIP text encoder heads responsible for processing negation and semantic nuances, thereby improving cross-modal alignment. Similarly, Cao et al. (2020) identified attention heads specialized for cross-modal interactions, integrating linguistic and visual cues for high-quality multimodal synthesis. Notably, it shares similarities with causal tracing, which can be applied once a layer has been broken down into distinct components using Representation Decomposition.

Main Findings and Gap. While CLIP and diffusion models are a great starting point for a case-study using representation decomposition, leveraging the inherent decomposability of transformers can be extended to understanding multimodal language models, and text-to-video models—an important gap that needs to be addressed.

2.5 General Task Vectors

General Task (or steering) vectors in language models are directional embeddings that, when added to specific layers, enhance model capabilities such as in-context learning and instruction following. To obtain these task vectors, one requires a well-annotated *supervised* probing dataset. Hendel et al. (2023a) discovered a task vector for compressing task demonstrations, while Zhang et al. (2024a) and Jiang et al. (2024a) leveraged instruction vectors to improve model adherence to user instructions and mitigate catastrophic forgetting. In multimodal models, task vectors facilitate controlled image generation and editing. Baumann et al. (2024) mapped text-embedding vectors to visual concepts for adjustable intensity, while Gandikota et al. (2025) fine-tuned low-rank matrices in UNet to create controllable concept vectors. Cohen et al. explored multiple task vectors in diffusion models, proposing a prompt-conditioned adaptation method to minimize interference.

Main Findings and Gap. While language models support both fine-tuning and zero-shot steering, multimodal models largely rely on fine-tuning. Advancing zero-shot steering for multimodal models remains a crucial research direction.

2.6 Sparse Autoencoders: A Special Class of Unsupervised Task Vectors

Sparse Autoencoders (SAEs, Yun et al. (2021)) offer an *unsupervised* approach to discovering conceptual representations in neural networks post-training. SAEs learn a dictionary of concepts such that any representation can be expressed as a linear combination of a *sparse* subset of these concepts. The SAE with an autoencoder architecture is trained to reconstruct its input while enforcing sparse activations. Once trained, neurons are interpreted based on their highest-activating inputs, forming a concept dictionary that maps concepts to vectors in representation space. These vectors can then be added to the model's residual stream to control attributes like safety and intensity in image generation. Due to their unsupervised nature, which minimizes the need for annotated examples for probing, SAEs have been applied extensively to LLMs to identify human interpretable directions for various concepts (e.g., refusal) in representation space (Cunningham et al., 2023). These directions can then be used to steer the language model (Marks et al., 2024) without the need of fine-tuning it. More recently, SAEs have been extended to vision-language models like CLIP (Daujotas, 2024; Rao et al., 2024; Lim et al., 2024) and audio transcription models like Whisper (Sadov, 2024). Despite their promise, SAEs face challenges such as feature absorption and splitting (Chanin et al., 2024), lack of robust evaluation metrics (Makelov et al., 2024) and underperformance compared to supervised methods for model control.

Main Findings and Gap. The effectiveness of SAEs as a control mechanism for multimodal models is still in its early stages and requires validation across a range of multimodal models, including the latest diffusion models and MLLMs.

2.7 Neuron-Level Descriptions

Neuron-level analysis methods aim to identify specific neurons that contribute to model predictions (Sajjad et al., 2022). In this section, we divide these methods into two main categories: gradient-based attribution, and activation-based analysis.[2] Gradient-based attribution methods analyze how neuron values influence model outputs by perturbing neuron activations and accumulating weight contributions based on corresponding gradients (Dai et al., 2021). In unimodal settings, Dai et al. (2021) detected fact-related neurons concentrated in the top layers of a pretrained language model, while Wang et al. (2022a) identified neurons for encoding hierarchical concepts in a CNN-based vision model. Extending this approach to multimodal settings, Schwettmann et al. (2023) identified "multimodal neurons" that transform visual representations into textual concepts via the model's residual stream. Activation-based analysis methods detect whether a neuron is activated when processing an input. These methods have been used to identify neurons specialized for specific tasks (Wang et al., 2022c) and multilingual understanding (Tang et al., 2024). Additionally, Voita et al. (2023) identified "dead" neurons that are never activated, revealing the sparsity of LLMs. In multimodal contexts, Goh et al. (2021) detected neurons encoding distinct visual features in non-generative models, while in generative VLMs, researchers have identified domain-specific neurons (Huo et al., 2024) and modality-specific neurons (Huang et al., 2024c). In diffusion models, Hintersdorf et al. (2024) identified memorization neurons by analyzing their out-of-distribution activations.

Main Findings and Gap. Neuron-level analysis adapts well to multimodal settings, but deeper neuron interactions remain underexplored, such as activation shifts in generative VLMs when adding visual input to identical text.

[2] Additional categories such as prediction probability changes and others are discussed in Appendix E.7.

3 Interpretability Methods Specific to Multimodal Models

In this section, we focus on ***mechanistic interpretability methods designed specifically for multimodal models***. These methods leverage architectural properties unique to multimodal systems such as cross-attention layers, or leverage the presence of a text-encoder to explain inner embeddings in human-understandable terms.

3.1 Text-Explanations of Embeddings

In Sec. 2.4, we leverage the representation decomposition property of transformers to identify key components in token representations. However, interpreting these components in human-understandable terms remains a challenge. For CLIP models, Gandelsman et al. (2024a) proposed TextSpan, which assigns textual descriptions to model components (e.g., attention heads) by identifying a text embedding that explains most of the variance in their outputs. The dataset for this task is *supervised* in nature. Expanding on this, Balasubramanian et al. (2024) introduced a scoring function to rank relevant textual descriptions across components. Concurrently, SpLiCE (Bhalla et al., 2024) mapped CLIP visual embeddings to sparse, interpretable concept combinations. Additionally, Parekh et al. (2024) employed dictionary learning to show that predefined concepts are semantically grounded in both vision and language. Together, these methods enhance the interpretability of internal embeddings in multimodal models by providing textual explanations.

Main Findings and Gap. Current text-based explanations of internal embeddings primarily focus on simple concepts (e.g., color, location). It remains unclear whether these methods can effectively map visual embeddings to more abstract concepts, such as physical laws. Moreover, their applicability beyond CLIP, particularly in text-to-image and video generation models, remains largely underexplored.

3.2 Network Dissection

Network Dissection (ND) (Bau et al., 2017), pioneered automated neuron interpretability in multimodal networks by establishing connections between individual neurons and human-understandable concepts. Different from the internal embedding methods (Sec. 3.1), ND compares neuron activations with groud-truth concept annotations in images. When a neuron's activation pattern consistently matches with a specific concept over a certain threshold, that concept is assigned as the neuron's interpretation (Oikarinen and Weng, 2023; Kalibhat et al., 2023). Moving beyond simple concept matching, MILAN (Hernandez et al., 2021) introduced a generative approach that produces natural language descriptions of neuron behavior based on highly activating images. DnD (Bai et al., 2024) then extend this work by first leveraging a generative VLM to describe highly activating images for each neuron and semantically combine these descriptions using an LLM.

Main Findings and Gap. The generalization of this method are constrained by their underlying multimodal architectures, e.g., CLIP. Moreover, while ND has proven effective for CNN-based vision models, its applicability to more advanced architectures, e.g., diffusion models, remains unexplored.

3.3 Cross-attention Based Interpretability

Cross-attention layers are crucial in multimodal models such as text-to-image diffusion models and generative VLMs, as they mediate interactions between image and text modalities. In generative models, studies have shown that cross-attention layers in UNet or DiT backbones play a critical role in linking an image's spatial layout to each word in the prompt (Tang et al., 2022). Building on this, Hertz et al. (2022) introduced a method for image editing via cross-attention control, enabling localized modifications, attribute amplification, and global changes while preserving image integrity. Similarly, Neo et al. (2024) identified memorization neurons within cross-attention layers, while Basu et al. (2024c) found that key concepts—such as artistic style, and factual knowledge—are concentrated in a small subset of these layers.

Main Findings and Gap. While the cross-attention mechanisms in U-Net-based diffusion models are well-studied for applications like image editing and compositionality, mechanistic analysis of cross-attention in diffusion transformers (DiTs) and generative VLMs for downstream applications remains an open research area.

3.4 Training Data Attribution Methods

Training data attribution identifies training examples crucial to a specific prediction or generation. Although well studied for non-generative vision models (Koh and Liang, 2020; Basu et al., 2021; Pruthi et al., 2020; Park et al., 2023), extending these methods to generative multimodal models (e.g., diffusion, multimodal language) remains challenging. Here, we highlight two categories of approaches specific to text-to-image diffusion models, with additional methods detailed in Appendix F.4.

(1) *Retrieval and Unlearning Based Methods.* A major challenge in training data attribution for diffusion models is the costly retraining needed for ground-truth influence and the adaptation of attribution methods due to time-step dependence. Wang et al. (2023b) evaluated retrieval-based attribution using image encoders (e.g., CLIP) as a baseline but did not incorporate diffusion model parameters. To address this, Wang et al. (2024b) introduced an unlearning-based approach, where generated images are "unlearned" by increasing their loss, creating an unlearned model. Attribution is then measured based on the deviation in training loss between the original and unlearned models, showing strong correlation with ground-truth attribution.

(2) *Gradient-Based Methods*, which are vital for data attribution in multimodal models, quantifying how training samples influence outputs via gradients. For diffusion models, adaptations include K-FAC (Mlodozeniec et al., 2024), which approximated the Generalized Gauss-Newton (GGN) matrix for scalable influence estimation, TRAK (Park et al., 2023), which modeled networks as kernel machines for improved attribution accuracy, and D-TRAK (Zheng et al., 2024b), which leveraged reverse diffusion and optimized gradient features for enhanced robustness. Additionally, DataInf (Kwon et al., 2024) bridged perturbation methods with influence function approximations. Collectively, these techniques refine gradient-based attribution by disentangling multimodal attribution patterns through targeted perturbations.

Main Findings and Gap. Multimodal data attribution is challenging due to the scale of heterogeneous pre-training data and complex model architectures, making retraining infeasible and inference slow. Efficient attribution methods and retraining-free evaluation techniques remain an open problem.

3.5 Feature Visualizations

In MMFMs, feature visualization techniques typically involve generating heatmaps of gradients or relevance scores over input images, providing an intuitive way to understand which features contribute to a model's final prediction. Grad-CAM (Selvaraju et al., 2017) firstly visualized a coarse localization map by tracking how gradients from a target concept flow back to the final prediction layer, highlighting key mage regions responsible for concept prediction. For both non-generative VLMs and MMFMs, this method has been employed to visualize grounding capabilities (Rajabi and Kosecka, 2024) and information flow in multimodal complex reasoning tasks (Zhang et al., 2024c). For diffusion models, Tang et al. (2022) aggregated cross-attention word–pixel scores within the denoising network to compute global attribution scores, thus showing how specific words in a text prompt influence different parts of a generated image. Instead of visualizing only the final generated images, Park et al. (2024) provided a more detailed view by visualizing regions of focus and the attention given to concepts from prompts at each denoising step.[3]

Main Findings and Gap. While feature visualization methods have been successfully applied to simple tasks such as image classification and visual question answering (VQA), their adaptation to more complex tasks—such as long-form image-to-text generation—remains underexplored.

4 Applications using Mechanistic Insights

In this section, we use the mechanistic insights from methods described in Sec. (2) and Sec. (3) for various downstream applications.

4.1 In-context Learning

Introduced in Sec. 2.5, Hendel et al. (2023b) and Liu et al. (2023c) establish that ICL in language models can be viewed through the lens of task vectors. Following these works, Huang et al. (2024a) characterizes multimodal task vectors as pairs of attention head activations and indices and applies those task vectors to generative VLMs in in-context learning settings to compress long prompts that would otherwise not fit in limited context length. Luo et al. (2024) further analyzes the transferability of task vectors from different modalities, which extends the application of task vectors.

[3]Additional details on relevance scores in are provided in Appendix F.5).

4.2 Model Editing

Building on Orgad et al. (2023), which modifies key and value matrices in cross-attention layers, Basu et al. (2024b) identifies and edits layers responsible for specific visual attributes. Using a brute-force approach, they intervene in cross-attention inputs and measure effects on generation, revealing that artistic styles, facts, and trademark objects are concentrated in a few layers, enabling efficient edits across text-to-image models. Basu et al. (2023) extends causal mediation analysis (Meng et al., 2022a) to text-to-image models, finding that, unlike LLMs, where causal layers vary, the first self-attention layer of the text encoder is the sole causal state, enabling targeted model edits. Basu et al. (2024a) applies causal tracing to Llava (Liu et al., 2023a) for factual VQA, modifying key layers to integrate long-tailed knowledge. While Pan et al. (2023) benchmarks language model editing techniques, these lack mechanistic insights. Compared to LLMs, large-batch and sequential editing remain underexplored in MMFMs.

4.3 Detecting and Mitigating Hallucinations

Dai et al. (2023) examines how image encodings (e.g., region, patch, grid) and loss functions impact hallucinations in contrastive and generative VLMs, proposing a lightweight fine-tuning method to mitigate them. Jiang et al. (2024b) finds that hallucinated objects have lower confidence when projected onto the output vocabulary, using this insight to develop a feature editing algorithm that removes them from captions. Jiang et al. (2024c) shows that real object tokens receive higher attention weights from visual tokens than hallucinated ones. Cohen et al. (2024) further analyzes visual-to-text information flow, offering insights for hallucination detection. Phukan et al. (2024) identifies logit lens limitations and introduces a similarity metric based on middle-layer embeddings to detect hallucinations. Overall, hallucination detection in MMFMs remains less explored compared to language models (Sakketou et al., 2022; Li et al., 2024b; Chen et al., 2024b; Cheng et al., 2023; Li et al., 2023c; Manakul et al., 2023). We also find that there is a lack of reliable benchmarking for hallucination detection methods for multimodal language models, when compared to language models.

4.4 Improving Safety

Early efforts to improve generative VLMs safety relied on fine-tuning (Zong et al., 2024), but recent work leverages mechanistic tools (Sec. 2, 3). Task vectors enhance safety by ablating harmful directions during inference (Wang et al., 2024a), while SAEs enforce sparsity to disentangle harmful features (Sharkey et al., 2022; Templeton et al., 2024). Xu et al. (2025) identifies hidden states crucial to safety mechanisms but find misalignment between modalities, proposing localized training to address it. In text-to-image models, SAEs help remove unwanted concepts (Cywiński and Deja, 2025; Ijishakin et al., 2024), and interpretable latent directions improve safe generations (Li et al., 2024a). For non-generative VLMs like CLIP, most work fine-tunes models for safety (Poppi et al., 2024), though interventional methods in (Basu et al., 2023; Gandelsman et al., 2024a) could help identify safety-related layers.

4.5 Improving Compositionality

Compositionality in text-to-image models refers to their ability to correctly represent object compositions, attributes, and relationships from a given prompt. Huang et al. (2023) introduces a benchmark to assess compositionality challenges in these models. LayoutGPT (Feng et al., 2024) leverages LLMs with few-shot learning to generate bounding boxes, guiding diffusion models via pixel-space loss. Grounded Compositional Generation (Phung et al., 2024) refines this by defining the loss in cross-attention space, improving performance. Similarly, Rassin et al. (2024) enhances attribute correspondence by aligning object-attention maps with adjectives. Beyond diffusion model modifications, some works address compositionality issues by improving text conditioning. Zarei et al. (2024) identifies erroneous attention in CLIP, where nouns misalign with adjectives, and proposes a projection layer to enhance attribute binding. Likewise, Zhuang et al. (2024) introduces a zero-shot method that adjusts object embeddings to strengthen relevant attribute associations while minimizing irrelevant ones.

5 Conclusion

Our survey reviews mechanistic understanding methods for MMFMs, including contrastive and generative VLMs and text-to-image diffusion models, with a focus on downstream applications. We introduce a novel taxonomy differentiating interpretability methods adapted from language models and those designed for multimodal models. Additionally, we compare mechanistic insights from language and multimodal models, identifying gaps in understanding and their impact on downstream applications.

6 Limitations

Our work has several limitations: (1) we mainly focus on the image-text multimodal model without considering other modalities such as video, time series, or 3D. (2) We don't contain the experimental analysis because of the lack of unified benchmarks. We will consider this in our future work. (3) We only focus on the transformer-based model or diffusion model, without considering novel model architecture such as MAMBA (Gu and Dao, 2023).

References

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

2023. Eliciting latent predictions from transformers with the tuned lens. *to appear*.

Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. 2022. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 21406–21415.

Abien Fred Agarap. 2019. Deep learning using rectified linear units (relu). *Preprint*, arXiv:1803.08375.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, and 23 others. 2024. Pixtral 12b. *Preprint*, arXiv:2410.07073.

Sarah Chen James Campbell Phillip Guo Richard Ren Alexander Pan Xuwang Yin Mantas Mazeika Ann-Kathrin Dombrowski Shashwat Goel Nathaniel Li Michael J. Byun Zifan Wang Alex Mallen Steven Basart Sanmi Koyejo Dawn Song Matt Fredrikson Zico Kolter Dan Hendrycks Andy Zou, Long Phan. 2023. Representation engineering: A top-down approach to ai transparency. *Preprint*, arXiv:2310.01405.

Omer Antverg and Yonatan Belinkov. 2021. On the pitfalls of analyzing individual neurons in language models. *arXiv preprint arXiv:2110.07483*.

Nicholas Bai, Rahul A Iyer, Tuomas Oikarinen, and Tsui-Wei Weng. 2024. Describe-and-dissect: Interpreting neurons in vision networks with language models. *arXiv preprint arXiv:2403.13771*.

Sriram Balasubramanian, Samyadeep Basu, and Soheil Feizi. 2024. Decomposing and interpreting image representations via text in vits beyond CLIP. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. 2024. What makes multimodal in-context learning work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1539–1550.

Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. 2024a. Understanding information storage and transfer in multi-modal large language models. *Preprint*, arXiv:2406.04236.

Samyadeep Basu, Vlad Morariu, Zichao Wang, Ryan Rossi, Cherry Zhao, Soheil Feizi, and Varun Manjunatha. 2025. On mechanistic circuits for extractive question-answering. *Preprint*, arXiv:2502.08059.

Samyadeep Basu, Philip Pope, and Soheil Feizi. 2021. Influence functions in deep learning are fragile. *Preprint*, arXiv:2006.14651.

Samyadeep Basu, Keivan Rezaei, Priyatham Kattakinda, Vlad I Morariu, Nanxuan Zhao, Ryan A Rossi, Varun Manjunatha, and Soheil Feizi. 2024b. On mechanistic knowledge localization in text-to-image generative models. In *Forty-first International Conference on Machine Learning*.

Samyadeep Basu, Nanxuan Zhao, Vlad I Morariu, Soheil Feizi, and Varun Manjunatha. 2023. Localizing and editing knowledge in text-to-image generative models. In *The Twelfth International Conference on Learning Representations*.

Samyadeep Basu, Nanxuan Zhao, Vlad I Morariu, Soheil Feizi, and Varun Manjunatha. 2024c. Localizing and editing knowledge in text-to-image generative models. In *The Twelfth International Conference on Learning Representations*.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549.

Stefan Andreas Baumann, Felix Krause, Michael Neumayr, Nick Stracke, Vincent Tao Hu, and Björn Ommer. 2024. Continuous, subject-specific attribute control in t2i models by identifying semantic directions. *Preprint*, arXiv:2403.17064.

Mohammad Beigi, Ying Shen, Runing Yang, Zihao Lin, Qifan Wang, Ankith Mohan, Jianfeng He, Ming Jin, Chang-Tien Lu, and Lifu Huang. 2024. Internalinspector i^2: Robust confidence estimation in llms through internal states. *Preprint*, arXiv:2406.12053.

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P Calmon, and Himabindu Lakkaraju. 2024. Interpreting clip with sparse linear concept embeddings (splice). *arXiv preprint arXiv:2402.10376*.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. *Preprint*, arXiv:2005.07310.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, and Joseph Bloom. 2024. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *Preprint*, arXiv:2409.14507.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791.

Qizhou Chen, Taolin Zhang, Chengyu Wang, Xiaofeng He, Dakan Wang, and Tingting Liu. 2024a. Attribution analysis meets model editing: Advancing knowledge correction in vision language models with visedit. *Preprint*, arXiv:2408.09916.

Shuo Chen, Zhen Han, Bailan He, Mark Buckley, Philip Torr, Volker Tresp, and Jindong Gu. 2023. Understanding and improving in-context learning on vision-language models. *arXiv preprint arXiv:2311.18021*, 1(2).

Xiang Chen, Duanzheng Song, Honghao Gui, Chenxi Wang, Ningyu Zhang, Yong Jiang, Fei Huang, Chengfei Lv, Dan Zhang, and Huajun Chen. 2024b. Factchd: Benchmarking fact-conflicting hallucination detection. *Preprint*, arXiv:2310.12086.

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, and Xipeng Qiu. 2023. Evaluating hallucinations in chinese large language models. *Preprint*, arXiv:2310.03368.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. Scaling instruction-finetuned language models.

Ido Cohen, Daniela Gottesman, Mor Geva, and Raja Giryes. 2024. Performance gap in entity knowledge extraction across modalities in vision language models. *Preprint*, arXiv:2412.14133.

Niv Cohen, Nicky Kriplani, Benjamin Feuer, Yuval Lemberg, and Chinmay Hegde. Multi-concept editing using task arithmetic.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023a. Towards automated circuit discovery for mechanistic interpretability. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023b. Towards automated circuit discovery for mechanistic interpretability. *Preprint*, arXiv:2304.14997.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *Preprint*, arXiv:2309.08600.

Bartosz Cywiński and Kamil Deja. 2025. Saeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders. *Preprint*, arXiv:2501.18052.

Adam Dahlgren Lindström, Johanna Björklund, Suna Bensch, and Frank Drewes. 2020. Probing multimodal embeddings for linguistic properties: the visual-semantic case. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 730–744, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2021. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*.

Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. 2023. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. *Preprint*, arXiv:2210.07688.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. 2019. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6309–6317.

Yunkai Dang, Kaichen Huang, Jiahao Huo, Yibo Yan, Sirui Huang, Dongrui Liu, Mengxi Gao, Jie Zhang, Chen Qian, Kun Wang, and 1 others. 2024. Explainable and interpretable multimodal large language models: A comprehensive survey. *arXiv preprint arXiv:2412.02104*.

Gytis Daujotas. 2024. Case study: Interpreting, manipulating, and controlling clip with sparse autoencoders. `https://www.lesswrong.com/posts/iYFuZo9BMvr6GgMs5/case-study-interpreting-manipulating-and-controlling-cl` Accessed: 2025-01-14.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, and 31 others. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. *Preprint*, arXiv:2409.17146.

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. 2020. Analyzing individual neurons in pre-trained language models. *arXiv preprint arXiv:2010.02695*.

Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer.

Benjamin L Edelman, Ezra Edelman, Surbhi Goel, Eran Malach, and Nikolaos Tsilivis. 2024. The evolution of statistical induction heads: In-context learning markov chains. *arXiv preprint arXiv:2402.11004*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, and 1 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. *Preprint*, arXiv:2403.03206.

Hao Fang, Yixiang Qiu, Hongyao Yu, Wenbo Yu, Jiawei Kong, Baoli Chong, Bin Chen, Xuan Wang, Shu-Tao Xia, and Ke Xu. 2024a. Privacy leakage on dnns: A survey of model inversion attacks and defenses. *arXiv preprint arXiv:2402.04013*.

Junfeng Fang, Zac Bi, Ruipeng Wang, Houcheng Jiang, Yuan Gao, Kun Wang, An Zhang, Jie Shi, Xiang Wang, and Tat-Seng Chua. Towards neuron attributions in multi-modal large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yingying Fang, Shuang Wu, Sheng Zhang, Chaoyan Huang, Tieyong Zeng, Xiaodan Xing, Simon Walsh, and Guang Yang. 2024b. Dynamic multimodal information bottleneck for multimodality classification. In *WACV*, pages 7681–7691. IEEE.

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2024. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36.

Jaden Fiotto-Kaufman, Alexander R Loftus, Eric Todd, Jannik Brinkmann, Caden Juang, Koyena Pal, Can Rager, Aaron Mueller, Samuel Marks, Arnab Sen Sharma, Francesca Lucchetti, Michael Ripa, Adam Belfki, Nikhil Prakash, Sumeet Multani, Carla Brodley, Arjun Guha, Jonathan Bell, Byron Wallace, and David Bau. 2024. Nnsight and ndif: Democratizing access to foundation model internals.

Deqing Fu, Tong Xiao, Rui Wang, Wang Zhu, Pengchuan Zhang, Guan Pang, Robin Jia, and Lawrence Chen. 2024. Tldr: Token-level detective reward model for large vision language models. *Preprint*, arXiv:2410.04734.

Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. 2024a. Interpreting clip's image representation via text-based decomposition. *Preprint*, arXiv:2310.05916.

Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. 2024b. Interpreting the second-order effects of neurons in clip. *arXiv preprint arXiv:2406.04341*.

Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. 2024c. Interpreting the second-order effects of neurons in clip. *Preprint*, arXiv:2406.04341.

Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. 2025. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, pages 172–188. Springer.

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. *Preprint*, arXiv:2406.04093.

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. 2022a. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. *arXiv preprint arXiv:2203.14680*.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2022b. Self-attention representations reveal the systematicity of semantic knowledge in bert. *arXiv preprint arXiv:2203.07442*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. *Preprint*, arXiv:2012.14913.

Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. 2021. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Danny Halawi, Jean-Stanislas Denain, and Jacob Steinhardt. 2024. Overthinking the truth: Understanding how language models process false demonstrations. *Preprint*, arXiv:2307.09476.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-attention attribution: Interpreting information interactions inside transformer. *Preprint*, arXiv:2004.11207.

Roee Hendel, Mor Geva, and Amir Globerson. 2023a. In-context learning creates task vectors. *Preprint*, arXiv:2310.15916.

Roee Hendel, Mor Geva, and Amir Globerson. 2023b. In-context learning creates task vectors. *arXiv preprint arXiv:2310.15916*.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. *Preprint*, arXiv:2106.09141.

Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. 2021. Natural language descriptions of deep visual features. In *International Conference on Learning Representations*.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control.

Dominik Hintersdorf, Lukas Struppek, Kristian Kersting, Adam Dziedzic, and Franziska Boenisch. 2024. Finding nemo: Localizing neurons responsible for memorization in diffusion models. *arXiv preprint arXiv:2406.02366*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *CoRR*, abs/2006.11239.

Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. 2020. exBERT: A Visual Analysis Tool to Explore Learned Representations in Transformer Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 187–196, Online. Association for Computational Linguistics.

Lijie Hu, Chenyang Ren, Huanyi Xie, Khouloud Saadi, Shu Yang, Jingfeng Zhang, and Di Wang. 2024. Dissecting misalignment of multimodal large language models via influence function. *Preprint*, arXiv:2411.11667.

Brandon Huang, Chancharik Mitra, Assaf Arbelle, Leonid Karlinsky, Trevor Darrell, and Roei Herzig. 2024a. Multimodal task vectors enable many-shot multimodal in-context learning. *arXiv preprint arXiv:2406.15334*.

Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. 2024b. Ravel: Evaluating interpretability methods on disentangling language model representations. *arXiv preprint arXiv:2402.17700*.

Kaichen Huang, Jiahao Huo, Yibo Yan, Kun Wang, Yutao Yue, and Xuming Hu. 2024c. Miner: Mining the underlying pattern of modality-specific neurons in multimodal large language models. *arXiv preprint arXiv:2410.04819*.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *arXiv preprint arXiv: 2307.06350*.

Jiahao Huo, Yibo Yan, Boren Hu, Yutao Yue, and Xuming Hu. 2024. Mmneuron: Discovering neuron-level domain-specific interpretation in multimodal large language model. *arXiv preprint arXiv:2406.11193*.

Ayodeji Ijishakin, Ming Liang Ang, Levente Baljer, Daniel Chee Hian Tan, Hugo Laurence Fry, Ahmed Abdulaal, Aengus Lynch, and James H. Cole. 2024. H-space sparse autoencoders. In *Neurips Safe Generative AI Workshop 2024*.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3543–3556. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Saurav Jha, Dong Gong, and Lina Yao. 2024. Clap4clip: Continual learning with probabilistic finetuning for vision-language models. *Preprint*, arXiv:2403.19137.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. *Preprint*, arXiv:2102.05918.

Gangwei Jiang, Caigao Jiang, Zhaoyi Li, Siqiao Xue, Jun Zhou, Linqi Song, Defu Lian, and Ying Wei. 2024a. Interpretable catastrophic forgetting of large language model fine-tuning via instruction vector. *arXiv preprint arXiv:2406.12227*.

Nick Jiang, Anish Kachinthaya, Suzie Petryk, and Yossi Gandelsman. 2024b. Interpreting and editing vision-language representations to mitigate hallucinations. *Preprint*, arXiv:2410.02762.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2024c. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. *Preprint*, arXiv:2411.16724.

Sonia Joseph. 2023. Vit prisma: A mechanistic interpretability library for vision transformers. `https://github.com/soniajoseph/vit-prisma`.

Curt Tigges Joseph Bloom and David Chanin. 2024. Saelens. `https://github.com/jbloomAus/SAELens`.

Neha Kalibhat, Shweta Bhardwaj, Bayan Bruss, Hamed Firooz, Maziar Sanjabi, and Soheil Feizi. 2023. Identifying interpretable subspaces in image representations. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Myeongseob Ko, Ming Jin, Chenguang Wang, and Ruoxi Jia. 2023. Practical membership inference attacks against large-scale multi-modal models: A pilot study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4871–4881.

Pang Wei Koh and Percy Liang. 2020. Understanding black-box predictions via influence functions. *Preprint*, arXiv:1703.04730.

Yongchan Kwon, Eric Wu, Kevin Wu, and James Zou. 2024. Datainf: Efficiently estimating data influence in lora-tuned llms and diffusion models. In *ICLR*. OpenReview.net.

João Lages. 2022. Diffusers-interpret. `https://github.com/JoaoLages/diffusers-interpret`.

Samuel Lavoie, Polina Kirichenko, Mark Ibrahim, Mahmoud Assran, Andrew Gordon Wilson, Aaron Courville, and Nicolas Ballas. 2024. Modeling caption diversity in contrastive vision-language pretraining. *Preprint*, arXiv:2405.00740.

Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. 2024a. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. *Preprint*, arXiv:2311.17216.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024b. The dawn after the dark: An empirical study on factuality hallucination in large language models. *Preprint*, arXiv:2401.03205.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023c. Halueval: A large-scale hallucination evaluation benchmark for large language models. *Preprint*, arXiv:2305.11747.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60.

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. *Preprint*, arXiv:2110.05208.

Paul Pu Liang, Yiwei Lyu, Gunjan Chhablani, Nihal Jain, Zihao Deng, Xingbo Wang, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2022. Multiviz: Towards visualizing and understanding multimodal models. *arXiv preprint arXiv:2207.00056*.

Hyesu Lim, Jinho Choi, Jaegul Choo, and Steffen Schneider. 2024. Sparse autoencoders reveal selective remapping of visual concepts during adaptation. *Preprint*, arXiv:2412.05276.

Aishan Liu, Yuguang Zhou, Xianglong Liu, Tianyuan Zhang, Siyuan Liang, Jiakai Wang, Yanjun Pu, Tianlin Li, Junqi Zhang, Wenbo Zhou, and 1 others. 2024a. Compromising embodied agents with contextual backdoor attacks. *arXiv preprint arXiv:2408.02882*.

Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Yu Cheng, and Wei Hu. 2024b. A survey of attacks on large vision-language models: Resources, advances, and future trends. *arXiv preprint arXiv:2407.07403*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Kevin Liu, Stephen Casper, Dylan Hadfield-Menell, and Jacob Andreas. 2023b. Cognitive dissonance: Why do language model outputs disagree with internal representations of truthfulness? *Preprint*, arXiv:2312.03729.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2023c. In-context vectors: Making in context learning more effective and controllable through latent space steering. *arXiv preprint arXiv:2311.06668*.

Xin Liu, Yichen Zhu, Yunshi Lan, Chao Yang, and Yu Qiao. 2024c. Safety of multimodal large language models on images and text. *arXiv preprint arXiv:2402.00357*.

Xinwei Liu, Xiaojun Jia, Yuan Xun, Siyuan Liang, and Xiaochun Cao. 2024d. Multimodal unlearnable examples: Protecting data against multimodal contrastive learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8024–8033.

Zhenhua Liu, Tong Zhu, Chuanyuan Tan, Haonan Lu, Bing Liu, and Wenliang Chen. 2024e. Probing language models for pre-training data detection. *Preprint*, arXiv:2406.01333.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Grace Luo, Trevor Darrell, and Amir Bar. 2024. Task vectors are cross-modal. *arXiv preprint arXiv:2410.22330*.

Yiwei Lyu, Paul Pu Liang, Zihao Deng, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. Dime: Fine-grained interpretations of multimodal models via disentangled local explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 455–467.

Aleksandar Makelov, George Lange, and Neel Nanda. 2024. Towards principled evaluations of sparse autoencoders for interpretability and control. *Preprint*, arXiv:2405.08366.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *Preprint*, arXiv:2303.08896.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2024. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *Preprint*, arXiv:2403.19647.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Chancharik Mitra, Brandon Huang, Tianning Chai, Zhiqiu Lin, Assaf Arbelle, Rogerio Feris, Leonid Karlinsky, Trevor Darrell, Deva Ramanan, and Roei Herzig. 2025. Sparse attention vectors: Generative multimodal model features are discriminative vision-language classifiers. *Preprint*, arXiv:2412.00142.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional chain-of-thought prompting for large multimodal models. In *CVPR*, pages 14420–14431. IEEE.

Bruno Mlodozeniec, Runa Eschenhagen, Juhan Bae, Alexander Immer, David Krueger, and Richard Turner. 2024. Influence functions for scalable data attribution in diffusion models. *Preprint*, arXiv:2410.13850.

W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

Neel Nanda and Joseph Bloom. 2022. Transformerlens. `https://github.com/TransformerLensOrg/TransformerLens`.

Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. 2023. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. 2024. Towards interpreting visual information processing in vision-language models. *Preprint*, arXiv:2410.07149.

Tuomas Oikarinen and Tsui-Wei Weng. 2023. CLIP-dissect: Automatic description of neuron representations in deep vision networks. In *The Eleventh International Conference on Learning Representations*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, and 1 others. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. 2023. Editing implicit assumptions in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7053–7061.

Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. 2023. Towards vision-language mechanistic interpretability: A causal tracing tool for blip. *Preprint*, arXiv:2308.14179.

Haowen Pan, Yixin Cao, Xiaozhi Wang, and Xun Yang. 2023. Finding and editing multi-modal neurons in pre-trained transformer. *arXiv preprint arXiv:2311.07470*.

Jayneel Parekh, Pegah Khayatan, Mustafa Shukor, Alasdair Newson, and Matthieu Cord. 2024. A concept-based explainability framework for large multimodal models. *arXiv preprint arXiv:2406.08074*.

Cheonbok Park, Inyoup Na, Yongjang Jo, Sungbok Shin, Jaehyo Yoo, Bum Chul Kwon, Jian Zhao, Hyungjong Noh, Yeonsoo Lee, and Jaegul Choo. 2019. Sanvis: Visual analytics for understanding self-attention networks. *Preprint*, arXiv:1909.09595.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8779–8788.

Ji-Hoon Park, Yeong-Joon Ju, and Seong-Whan Lee. 2024. Explaining generative diffusion models via visual analysis for interpretable decision-making process. *Expert Systems with Applications*, 248:123231.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. 2023. TRAK: attributing model behavior at scale. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 27074–27113. PMLR.

Judea Pearl. 2001. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, page 411–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Judea Pearl. 2014. Interpretation and identification of causal mediation. *Psychological methods*, 19.

Wenshuo Peng, Kaipeng Zhang, Yue Yang, Hao Zhang, and Yu Qiao. 2024. Data adaptive traceback for vision-language foundation models in image classification. In *AAAI*, pages 4506–4514. AAAI Press.

Anirudh Phukan, Divyansh, Harshit Kumar Morj, Vaishnavi, Apoorv Saxena, and Koustava Goswami. 2024. Beyond logit lens: Contextual embeddings for robust hallucination detection grounding in vlms. *Preprint*, arXiv:2411.19187.

Quynh Phung, Songwei Ge, and Jia-Bin Huang. 2024. Grounded text-to-image synthesis with attention refocusing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7932–7942.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *Preprint*, arXiv:2307.01952.

Samuele Poppi, Tobia Poppi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Safe-clip: Removing nsfw concepts from vision-and-language models. *Preprint*, arXiv:2311.16254.

Vidya Prasad, Chen Zhu-Tian, Anna Vilanova, Hanspeter Pfister, Nicola Pezzotti, and Hendrik Strobelt. 2023. Unraveling the temporal dynamics of the unet in diffusion models. *Preprint*, arXiv:2312.14965.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. 2019. Learning to deceive with attention-based explanations. *arXiv preprint arXiv:1909.07913*.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating training data influence by tracing gradient descent. In *NeurIPS*.

Shuhan Qi, Zhengying Cao, Jun Rao, Lei Wang, Jing Xiao, and Xuan Wang. 2023. What is the limitation of multimodal llms? a deeper look into multimodal llms through prompt probing. *Information Processing & Management*, 60(6):103510.

Libo Qin, Qiguang Chen, Hao Fei, Zhi Chen, Min Li, and Wanxiang Che. 2024. What factors affect multimodal in-context learning? an in-depth exploration. *arXiv preprint arXiv:2410.20482*.

Luyu Qiu, Yi Yang, Caleb Chen Cao, Yueyuan Zheng, Hilary Hei Ting Ngai, Janet Hui-wen Hsiao, and Lei Chen. 2022. Generating perturbation-based explanations with robustness to out-of-distribution data. In *WWW*, pages 3594–3605. ACM.

Vincent Quantmeyer, Pablo Mosteiro, and Albert Gatt. 2024. How and where does clip process negation? *Preprint*, arXiv:2407.10488.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Navid Rajabi and Jana Kosecka. 2024. Q-groundcam: Quantifying grounding in vision language models via gradcam. *arXiv preprint arXiv:2404.19128*.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. 2024. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *Preprint*, arXiv:2407.14435.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *Preprint*, arXiv:2204.06125.

Jinmeng Rao, Song Gao, Gengchen Mai, and Krzysztof Janowicz. 2023. Building privacy-preserving and secure geospatial artificial intelligence foundation models (vision paper). In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pages 1–4.

Sukrut Rao, Sweta Mahajan, Moritz Böhle, and Bernt Schiele. 2024. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In *European Conference on Computer Vision*, pages 444–461.

Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. 2024. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, 36.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Marko Robnik-Šikonja and Igor Kononenko. 2008. Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, 20(5):589–600.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. *Preprint*, arXiv:2112.10752.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597.

Konstantine Sadov. 2024. Feature discovery in audio models: A whisper case study. `https://builders.mozilla.org/insider-whisper/`. Accessed: 2025-01-14.

Hassan Sajjad, Nadir Durrani, and Fahim Dalvi. 2022. Neuron-level interpretation of deep nlp models: A survey. *Transactions of the Association for Computational Linguistics*, 10:1285–1303.

Flora Sakketou, Joan Plepi, Riccardo Cervero, Henri Jacques Geiss, Paolo Rosso, and Lucie Flek. 2022. FACTOID: A new dataset for identifying misinformation spreaders and political bias. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3231–3241, Marseille, France. European Language Resources Association.

Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. 2022. Are vision-language transformers learning multimodal representations? a probing perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11248–11257.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. 2022. Confident adaptive language modeling. *Preprint*, arXiv:2207.07061.

Sarah Schwettmann, Neil Chowdhury, Samuel Klein, David Bau, and Antonio Torralba. 2023. Multimodal neurons in pretrained text-only transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2862–2867.

Sarah Schwettmann, Tamar Shaham, Joanna Materzynska, Neil Chowdhury, Shuang Li, Jacob Andreas, David Bau, and Antonio Torralba. 2024. Find: A function description benchmark for evaluating interpretability methods. *Advances in Neural Information Processing Systems*, 36.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Lee Sharkey, Dan Braun, and Beren Millidge. 2022. Taking features out of superposition with sparse autoencoders. In *AI Alignment Forum*, volume 6, pages 12–13.

Gabriela Ben Melech Stan, Raanan Yehezkel Rohekar, Yaniv Gurwicz, Matthew Lyle Olson, Anahita Bhiwandiwalla, Estelle Aflalo, Chenfei Wu, Nan Duan, Shao-Yen Tseng, and Vasudev Lal. 2024. Lvlm-intrepret: An interpretability tool for large vision-language models. *arXiv preprint arXiv:2404.03118*.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. *Preprint*, arXiv:2305.15054.

Shilin Sun, Wenbin An, Feng Tian, Fang Nan, Qidong Liu, Jun Liu, Nazaraf Shah, and Ping Chen. 2024. A review of multimodal explainable artificial intelligence: Past, present and future. *arXiv preprint arXiv:2412.14056*.

Viacheslav Surkov, Chris Wendler, Mikhail Terekhov, Justin Deschenaux, Robert West, and Caglar Gulcehre. 2024. Unpacking sdxl turbo: Interpreting text-to-image models with sparse autoencoders. *Preprint*, arXiv:2410.22366.

Aaquib Syed, Can Rager, and Arthur Conmy. 2023. Attribution patching outperforms automated circuit discovery. *Preprint*, arXiv:2310.10348.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. 2022. What the daam: Interpreting stable diffusion using cross attention. *arXiv preprint arXiv:2210.04885*.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. *arXiv preprint arXiv:2402.16438*.

Mingxu Tao, Quzhe Huang, Kun Xu, Liwei Chen, Yansong Feng, and Dongyan Zhao. 2024. Probing multimodal large language models for global and local semantic representations. *Preprint*, arXiv:2402.17304.

Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *Preprint*, arXiv:2405.09818.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, and 1 others. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. transformer circuits thread.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Hannes Thurnherr and Jérémy Scheurer. 2024. Tracrbench: Generating interpretability testbeds with large language models. *arXiv preprint arXiv:2409.13714*.

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. 2020. Intrinsic probing through dimension selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 197–216, Online. Association for Computational Linguistics.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Preprint*, arXiv:2106.13884.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across nlp tasks. *arXiv preprint arXiv:1909.11218*.

Jesse Vig. 2019. Visualizing attention in transformer-based language models. *arXiv preprint arXiv:1904.02679*.

Theia Vogel. 2024. repeng.

Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. 2023. Neurons in large language models: Dead, n-gram, positional. *arXiv preprint arXiv:2309.04827*.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *Preprint*, arXiv:1905.09418.

Andong Wang, Wei-Ning Lee, and Xiaojuan Qi. 2022a. Hint: Hierarchical neuron concept explainer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10254–10264.

Han Wang, Gang Wang, and Huan Zhang. 2024a. Steering away from harm: An adaptive approach to defending vision language model against jailbreaks. *Preprint*, arXiv:2411.16721.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022b. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *Preprint*, arXiv:2211.00593.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023a. Label words are anchors: An information flow perspective for understanding in-context learning. *arXiv preprint arXiv:2305.14160*.

Sheng-Yu Wang, Alexei A. Efros, Jun-Yan Zhu, and Richard Zhang. 2023b. Evaluating data attribution for text-to-image models. In *ICCV*.

Sheng-Yu Wang, Alexei A. Efros, Jun-Yan Zhu, and Richard Zhang. 2023c. Evaluating data attribution for text-to-image models. In *ICCV*, pages 7158–7169. IEEE.

Sheng-Yu Wang, Aaron Hertzmann, Alexei A. Efros, Jun-Yan Zhu, and Richard Zhang. 2024b. Data attribution for text-to-image models by unlearning synthesized images. *Preprint*, arXiv:2406.09408.

Xiaozhi Wang, Kaiyue Wen, Zhengyan Zhang, Lei Hou, Zhiyuan Liu, and Juanzi Li. 2022c. Finding skill neurons in pre-trained transformer-based language models. *arXiv preprint arXiv:2211.07349*.

Ying Wang, Tim G. J. Rudner, and Andrew Gordon Wilson. 2023d. Visual explanations of image-text representations via multi-modal information bottleneck attribution. In *NeurIPS*.

Youze Wang, Wenbo Hu, Yinpeng Dong, Hanwang Zhang, Hang Su, and Richang Hong. 2024c. Exploring transferability of multimodal adversarial samples for vision-language pre-training models with contrastive learning. *Preprint*, arXiv:2308.12636.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

Zhengxuan Wu, Atticus Geiger, Aryaman Arora, Jing Huang, Zheng Wang, Noah D. Goodman, Christopher D. Manning, and Christopher Potts. 2024. pyvene: A library for understanding and improving PyTorch models via interventions.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. 2024a. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*.

Tong Xie, Haoyu Li, Andrew Bai, and Cho-Jui Hsieh. 2024b. Data attribution for diffusion models: Timestep-induced bias in influence estimation. *Trans. Mach. Learn. Res.*, 2024.

Ji Xin, Raphael Tang, Jaejun Lee, Yaoliang Yu, and Jimmy Lin. 2020. DeeBERT: Dynamic early exiting for accelerating BERT inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2246–2251, Online. Association for Computational Linguistics.

Shicheng Xu, Liang Pang, Yunchang Zhu, Huawei Shen, and Xueqi Cheng. 2025. Cross-modal safety mechanism transfer in large vision-language models. In *The Thirteenth International Conference on Learning Representations*.

Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. 2024. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. *arXiv preprint arXiv:2410.02761*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024a. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Fan Yang, Yihao Huang, Kailong Wang, Ling Shi, Geguang Pu, Yang Liu, and Haoyu Wang. 2024b. Efficient and effective universal adversarial attack against vision-language pre-training models. *arXiv preprint arXiv:2410.11639*.

Xikang Yang, Xuehai Tang, Fuqing Zhu, Jizhong Han, and Songlin Hu. 2024c. Enhancing cross-prompt transferability in vision-language models through contextual injection of target tokens. *Preprint*, arXiv:2406.13294.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *Preprint*, arXiv:2111.07783.

Hao Yin, Guangzong Si, and Zilei Wang. 2024. Unraveling the shift of visual information flow in MLLMs: From phased interaction to efficient inference.

Zeping Yu and Sophia Ananiadou. 2024a. Interpreting arithmetic mechanism in large language models through comparative neuron analysis. *arXiv preprint arXiv:2409.14144*.

Zeping Yu and Sophia Ananiadou. 2024b. Neuron-level knowledge attribution in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3267–3280.

Zeping Yu and Sophia Ananiadou. 2024c. Understanding multimodal llms: the mechanistic interpretability of llava in visual question answering. *Preprint*, arXiv:2411.10950.

Zeping Yu and Sophia Ananiadou. 2024d. Understanding multimodal llms: the mechanistic interpretability of llava in visual question answering. *arXiv preprint arXiv:2411.10950*.

Zeyu Yun, Yubei Chen, Bruno Olshausen, and Yann LeCun. 2021. Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 1–10, Online. Association for Computational Linguistics.

Arman Zarei, Keivan Rezaei, Samyadeep Basu, Mehrdad Saberi, Mazda Moayeri, Priyatham Kattakinda, and Soheil Feizi. 2024. Understanding and mitigating compositional issues in text-to-image generative models. *arXiv preprint arXiv:2406.07844*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *Preprint*, arXiv:2303.15343.

Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. 2024a. Tell your model where to attend: Post-hoc attention steering for llms. *Preprint*, arXiv:2311.02262.

Shaolei Zhang, Tian Yu, and Yang Feng. 2024b. Truthx: Alleviating hallucinations by editing large language models in truthful space. *Preprint*, arXiv:2402.17811.

Xiaofeng Zhang, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. 2024c. From redundancy to relevance: Enhancing explainability in multimodal large language models. *arXiv preprint arXiv:2406.06579*.

Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. 2023. Inversion-based style transfer with diffusion models. *Preprint*, arXiv:2211.13203.

Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Gould. 2024a. The first to know: How token distributions reveal hidden knowledge in large vision-language models? *Preprint*, arXiv:2403.09037.

Tianyi Zhao, Liangliang Zhang, Yao Ma, and Lu Cheng. 2024b. A survey on safe multi-modal learning systems. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6655–6665.

Haonan Zheng, Wen Jiang, Xinyang Deng, and Wenrui Li. 2024a. Sample-agnostic adversarial perturbation for vision-language pre-training models. In *ACM Multimedia*, pages 9749–9758. ACM.

Xiaosen Zheng, Tianyu Pang, Chao Du, Jing Jiang, and Min Lin. 2024b. Intriguing properties of data attribution on diffusion models. In *ICLR*. OpenReview.net.

Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. 2024. Visual in-context learning for large vision-language models. *arXiv preprint arXiv:2402.11574*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *Preprint*, arXiv:2304.10592.

Chenyi Zhuang, Ying Hu, and Pan Gao. 2024. Magnet: We never know how text-to-image diffusion models work, until we learn how vision-language models function. *arXiv preprint arXiv:2409.19967*.

Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. 2017. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595*.

Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. 2024. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. *Preprint*, arXiv:2402.02207.

A Comparison with Previous Surveys

Recently, Dang et al. (2024) provides a broad overview of interpretability methods for MMFMs across data, model architecture, and training paradigms. Another concurrent work (Sun et al., 2024) reviews the multimodal interpretability methods from a historical view, covering works from 2000 to 2025. While insightful, our work differs from theirs in both focus and scope. To be specific, our work examines how established LLM interpretability techniques adapt to various multimodal models, analyzing key differences between unimodal and multimodal systems in techniques, applications, and findings.

B Taxonomy Details

In our survey, we present an easy-to-read taxonomy that categorizes mechanistic interpretability techniques along three dimensions: (i) Dimension 1 categorizes whether the technique has been used for language models (Sec.2) or is specifically designed for multimodal models (Sec.3); (ii)) Dimension 2 provides a view of the mechanistic insights across various multimodal model families including non-generative VLMs (e.g., CLIP), text-to-image models (e.g., Stable-Diffusion) and multimodal language models (e.g., LLaVa). We describe the architectures studied in our paper in Sec.(C) and discuss their relevant mechanistic insights in Sec.(2) and Sec.(3). (iii) Dimension 3 links insights from these mechanistic methods to downstream practical applications (Sec.4). The taxonomy is visualized in Figure 1. In particular, the distribution of insights and applications are in-line in Sec. (2, 3, 4).

We believe this simple categorization will help readers (i) understand the gaps between unimodal language models and multimodal models in terms of mechanistic insights and applications, and (ii) identify the multimodal models where mechanistic interpretability (and their applications) is underexplored.

C Additional Details on Model Architectures

In this section, we introduce three main categories of multimodal models covered by our survey, including (i) Contrastive (i.e., Non-Generative) Vision-Language Models, Generative Vision-Language Models, and Text-to-image Diffusion Models. We choose these three families as they encompass the majority of the state-of-the-art architectures used by the community currently.

C.1 Non-Generative Vision-Language Models

One non-generative vision-language model (e.g., CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), FILIP (Yao et al., 2021), SigCLIP (Zhai et al., 2023), DeCLIP (Li et al., 2022) and LLIP (Lavoie et al., 2024)) usually contains one language-model-based text encoder and one vision-model-based vision encoder. These models are particularly suited for real-world applications such as text-guided image retrieval, image-guided text retrieval and zero-shot image classification.

C.2 Text-to-Image Diffusion Models

State-of-the-art text-guided image generation models are primarily based on the diffusion objective (Rombach et al., 2022; Ho et al., 2020), which predicts the noise that was added during the forward diffusion process, allowing it to learn how to gradually denoise random Gaussian noise back into a clean image during the reverse diffusion process. One diffusion model often contains a text encoder (e.g., CLIP) and a CNN-based U-Net (Ronneberger et al., 2015) for denoising to generate images. Early variants of text-to-image generative models with this objective include Stable-Diffusion-1 (Rombach et al., 2022) (which perform the diffusion process in a compressed latent space) and Dalle-2 (Ramesh et al., 2022) (which perform the diffusion process in the image space instead of a compressed latent space). In recent times, SD-XL (Podell et al., 2023) improves on the early Stable-Diffusion variants by using a larger denoising UNet and an improved conditioning (e.g., text or image) mechanism. More recent models such as Stable-Diffusion-3 (Esser et al., 2024) obtain stronger image generation results than previous Stable-Diffusion variants by (i) using a rectified flow formulation, (ii) scalable transformer architecture as the diffusion backbone and (iii) using an ensemble of strong text-encoders (e.g., T5 (Raffel et al., 2020; Chung et al., 2022)). Beyond image generation, in terms of downstream applications, text-to-image models can also be applied for image editing (Hertz et al., 2022), and style transfer (Zhang et al., 2023).

C.3 Generative Vision-Language Models

In our paper, we investigate the most common generative VLMs which are developed by connecting a vision encoder (e.g., CLIP) to a large language model through a bridge module. This bridge module (e.g., a few MLP layers (Liu et al., 2023a) or a Q-former (Li et al., 2023a)) is then trained on large-scale image-text pairs. Frozen (Tsimpoukelli et al., 2021) is one of the first works to take advantage of a large language model in image understanding tasks (e.g., few-shot learning). Follow-up works such as MiniGpt (Zhu et al., 2023), BLIP variants (Li et al., 2023b) and LLava (Liu et al., 2023a) improved on Frozen by modifying the scale and type of the training data, as well as the underlying architecture. In recent times, much focus has been geared toward curating high-quality image-text pairs encompassing various vision-language tasks. Qwen (Yang et al., 2024a), Pixtral (Agrawal et al., 2024) and Molmo (Deitke et al., 2024) are some of the recent multimodal language models focusing on high-quality image-text curated data. Multimodal language models have various real-world applications, such as VQA, and image captioning.

D More Definitions

We define a type of interpretability method as "supervised" if we need to have a labeled dataset to analyze it, otherwise, it is "unsupervised".

In the following sections, we also classify the papers in each type of method from the following perspective: (1) the interpretability aspect - what the method aims to interpret, e.g., data influence, fine-tuning, information flow, knowledge localization, and component contribution. (2) The analyzed component of a model, e.g., emebddings, layers (MLP, self attention, cross attention), or more fine-grained neurons. The illustration of model components is shown in Figure 2. (3) Applications: the downstream applications that are inspired by the insights of this method. Note, this is different from the task column in Table 7 and 10 which represents the task each paper they use to conduct interpretability analysis.

E Additional Details on Section 2

We add additional details about the interpretability methods adapted from LLM models.

E.1 Additional Details on Linear Probing

The linear probing method is very flexible and be applied for various interpretability purposes, and model components, and can also inspire various downstream tasks. We summarize all the papers of linear probing in Table 1.

E.2 Additional Details on Logit Lens

As Linear Probing, the Logit Lens is another flexible method. We summarize all the related papers in Table 2.

E.3 Additional Details on Causal Tracing

While causal tracing helps to identify individual "causal" components for a particular task, it does not automatically lead to the extraction of a subgraph of the underlying computational graph of a model which is "causal" for a task. In this regard, there has been a range of works in language modeling to extract task-specific circuits (Syed et al., 2023; Wang et al., 2022b; Conmy et al., 2023b). However, extending these methods to obtain task-specific circuits is still an open problem for MMFMs.

E.4 Additional Details on Representation Decomposition

In transformer-based LLMs, the concept of representation decomposition pertains to the analysis of the model's internal mechanisms, specifically dissecting individual transformer layers to core meaningful components, which aims at understanding the inner process of transformers. In unimodal LLMs, research has mainly decomposed the architecture and representation of a model's layer into two principal components: the attention mechanism and the multi-layer perceptron (MLP) layer. Intensive research efforts have focused on analyzing these components to understand their individual contributions to the model's decision-making process. Studies find that while attention should not be directly equated with explanation (Pruthi et al., 2019; Jain and Wallace, 2019; Wiegreffe and Pinter, 2019), it provides significant insights into the model's operational behavior and helps in error diagnosis and hypothesis development (Park et al., 2019; Voita et al., 2019; Vig, 2019; Hoover et al., 2020; Vashishth et al., 2019). Furthermore, concurrently, research has shown that Feed-Forward Networks (FFNs) within the Transformer MLP layer, functioning as key-value memories, encode and retrieve factual and semantic knowledge (Geva et al., 2021). Experimental studies have established a direct correlation between modifications in FFN output distributions and subsequent token probabilities, suggesting that the model's output is crafted

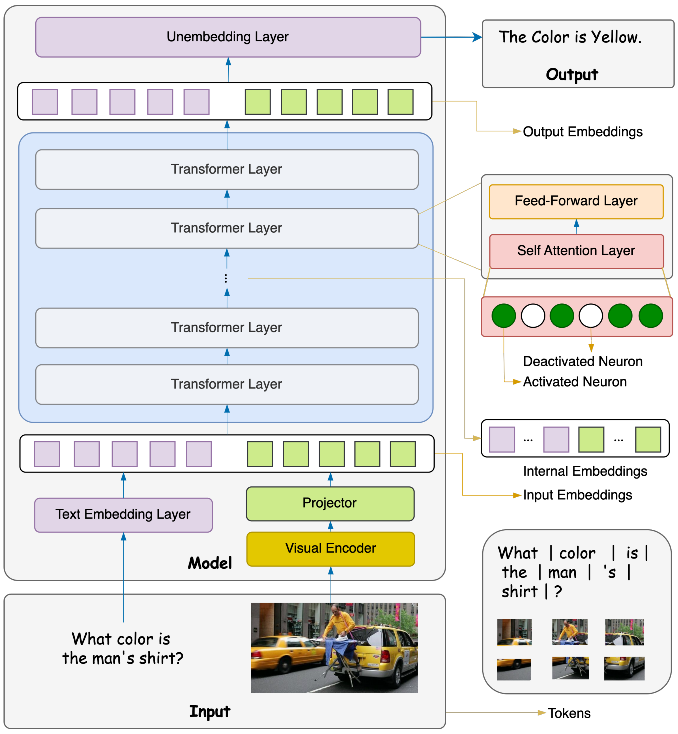
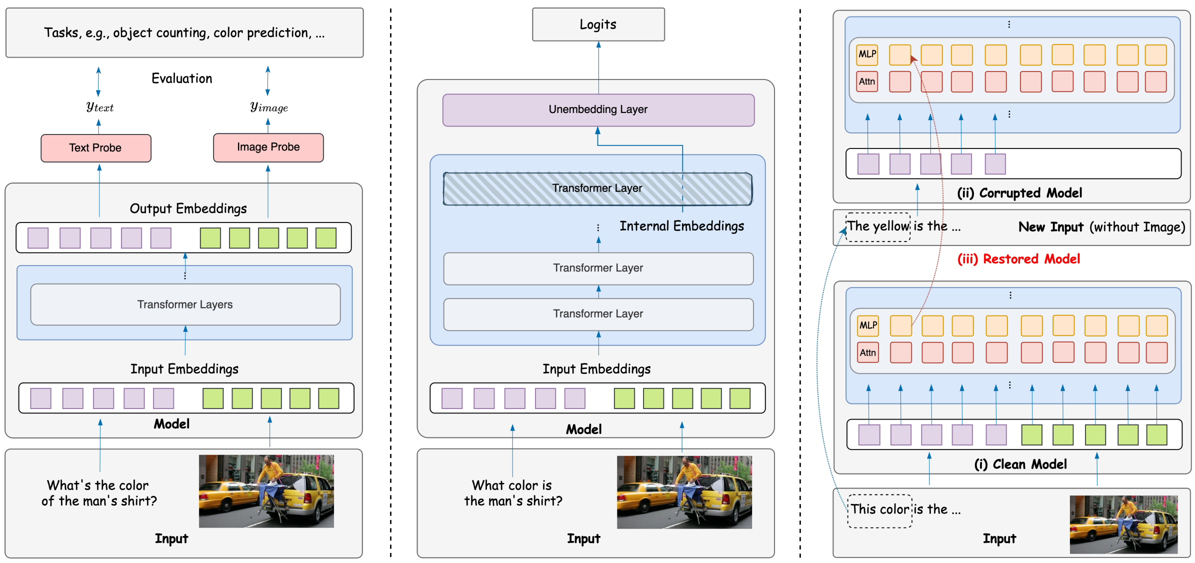


Figure 2: The illustration of model components. Take the transformer-based generative vision-language model as an example.



(a) Linear Probing

(b) Logit Lens

(c) Causal Tracing

Figure 3: The illustrations of interpretability methods: (a) Linear Probing, (b) Logit Lens, and (c) Causal Tracing.

through cumulative updates from each layer (Geva et al., 2022b). This core property serves as the foundation for identifying language model circuits associated with specific tasks in (Syed et al., 2023; Wang et al., 2022c; Conmy et al., 2023a).

E.5 Additional Details on General Task Vectors

We summarize all the related papers in Table 4.

E.6 Additional Details on Sparse Autoencoders

An SAE is typically a two-layer MLP of the form $SAE(x) = Dec(Act(Enc(x)))$ where x is the input feature. The encoder (Enc) and the decoder (Dec) layers are simple linear layers and the activation function (Act) is a design choice and can be a simple ReLU (Agarap, 2019), Top K (Gao

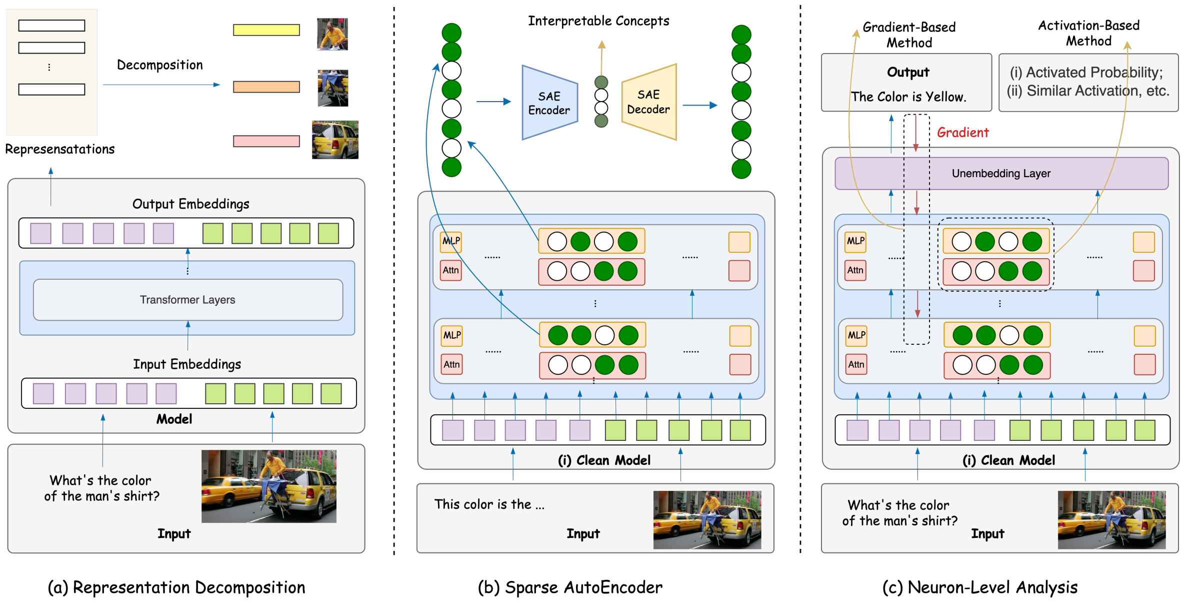


Figure 4: The illustrations of interpretability methods: (a) Representation Decomposition, (b) Sparse AutoEncoder, and (c) Neuron-level Analysis.

Paper	Interpretability Aspect	Analyzed Component	Application
(Tao et al., 2024)	Information flow	Layers	Visual-language entailment
(Torroba Hennigen et al., 2020)	Knowledge localization	Neurons	Linguistic understanding
(Dahlgren Lindström et al., 2020)	Knowledge localization	Image-text embedding	Image-caption alignment
(Dai et al., 2023)	Component contribution	Image encoding	Object hallucination
(Cao et al., 2020)	Information flow	Cross modal interaction	V+L benchmark
(Salin et al., 2022)	Component contribution	Layers	Multimodal understanding
(Qi et al., 2023)	Data influence	Prompt	Prompt optimization

Table 1: Additional Details on Linear Probing Papers

et al., 2024), JumpReLU (Rajamanoharan et al., 2024), and so on. The SAE is trained to reconstruct its own input, with the constraint that the activations should be sparse. Once trained, the neurons in the activation layer are assigned interpretations based on the highest activating input samples for the specific neuron in question. This results in a concept dictionary where concepts are mapped to directions (i.e., *vectors*) in representation space. These vectors can be added to the residual stream of the model to potentially control various facets such as the safety and intensity of various attributes in image generation models.

Overall, all the papers on Sparse Autoencoders analysis aim to interpret where the knowledge is stored in the model by analyzing the layers (as well as neurons). The inspired application is only the model steering.

E.7 Additional Details on Neuron-Level Analysis

There are different definitions of neurons in deep neural networks. We define $\mathbf{x}$ as the input embeddings, and $\mathbf{h_i}$ as the hidden states of the i-th layer's output. A model layer multiplies the hidden states with parameter M_i followed by an activation function $\mathbf{a} = f(xM_i^{\intercal})$. Some studies define the activation a_j, which is the j-th element of $\mathbf{a}$ as the neuron (Dai et al., 2021). While other works (Dalvi et al., 2019; Durrani et al., 2020; Antverg and Belinkov, 2021) define the dimensions in output representation as a neuron. For consistency, in our survey, we follow the most widely used definition to define an element m_j of a layer's parameter M as the neuron.

Prediction Probability Changes methods usually change the neuron output value, and analyze its influence on the final prediction. Yu and Ananiadou (2024b) quantifies the importance level of

Paper	Interpretability Aspect	Analysed Component	Application
(Phukan et al., 2024)	Data Influence	Hidden states	Improving VQA Performance
(Jiang et al., 2024c)	Information flow	Attention heads	Object hallucination
(Huo et al., 2024)	Knowledge localization	Neurons	-
(Zhao et al., 2024a)	Information flow	Hidden states	Controllable generation

Table 2: Additional Details on Logit Len Papers

Paper	Interpretability Aspect	Analysed Component	Application
(Basu et al., 2023)	Knowledge Localization	Self-attention	Model Editing
(Basu et al., 2024c)	Knowledge Localization	Cross-attention	Model Editing
(Basu et al., 2024a)	Knowledge Localization, Flow	MLP	Model Editing
(Yu and Ananiadou, 2024c)	Knowledge Localization	Self-attention	-
(Palit et al., 2023)	Knowledge Localization	Self-attention	-

Table 3: Additional Details on Causal Tracing Papers

a neuron by calculating the difference of the log of the probabilities by giving and without giving the neuron value. In this way, this paper finds that both attention and FFN layer store knowledge. Besides, all important neurons directly contributing to knowledge prediction are in deep layers. Yu and Ananiadou (2024a) utilizes the same method to find that features are enhanced in shallow FFN layers and neurons in deep layers are used to enhance prediction. Following a similar strategy, Yu and Ananiadou (2024d) finds important attention heads for handling VQA tasks.

Attribution Method is to project the internal hidden representation into output space to analyze each neuron's contribution to the final prediction (Geva et al., 2022a). In the multimodal domain, Pan et al. (2023) projects the activation of one neuron into output space to quantify the importance of one neuron to the final prediction and identify multimodal neurons. Fang et al. utilizes this method to find the semantic knowledge neurons and some interesting properties such as cross-modal invariance and semantic sensitivity.

Other Method covers many different types of neuron-level analysis methods. For example, instead of directly analyzing the first-order effect, which is the logits of each neuron, Gandelsman et al. (2024b) analyzes the accumulation of information of a neuron after the attention head. A new method to analyze information flow. Focus on the contribution of neurons to the output representation.

E.8 Summary

Overall, we find that the core principles of popular LLM-based mechanistic interpretability methods can be extended to multimodal models without complex modification. However, extracting meaningful mechanistic insights from these models often requires carefully tailored adaptations. In Table 7 - Appendix, we provide an overall comprehensive listing and analysis of all the papers discussed in this section. This table includes more detailed information on the datasets utilized, the models employed, and the specific tasks they conduct analysis experiments on. Note, that the "task" is different from "application" in the tables of each method, which is inspired by interpretability findings.

F Additional Details on Section 3

F.1 Additional Details on Text-Explanations of Internal Embeddings

All the text-explanations of internal embedding papers aim to interpret where knowledge is stored in the model. We summarize the papers in Table 6.

F.2 Additional Details on Network Dissect

Network Dissect mainly aims to localizing knowledge storage in network or visual representations. We summarize the related papers in Table 8.

F.3 Additional Details on Cross-attention Interpretability

We summarized the related papers in Table 9.

Paper	Interpretability Aspect	Analyzed Component	Application
(Baumann et al., 2024)	Fine-tuning	Layers	Continuous Image Editing
(Gandikota et al., 2025)	Fine-tuning	LoRA Layers	Continuous Image Editing
(Cohen et al.)	Knowledge Localization	Layers	Model Editing

Table 4: Additional Details on General Task Vectors Papers

Paper	Interpretability Aspect	Analyzed Component	Application
(Daujotas, 2024)	Knowledge Localization	Layers,Neurons	Model Steering
(Rao et al., 2024)	Knowledge Localization	Layers,Neurons	Model Steering
(Lim et al., 2024)	Knowledge Localization	Layers,Neurons	Model Steering
(Surkov et al., 2024)	Knowledge Localization	Layers,Neurons	Model Steering
(Sadov, 2024)	Knowledge Localization	Layers, Neurons	Model Steering

Table 5: **Additional Details on Sparse-Autoencoders**

F.4 Additional Details on Training Data Attribution

Training Dynamics-Based Methods These methods analyze how model parameters and predictions evolve during training to determine the influence of specific data points, thereby revealing how models learn from and prioritize instances. However, applying them to multimodal or generative models—like diffusion models—poses challenges. For instance, Training Data Influence (TracIn) (Pruthi et al., 2020) can suffer from "timestep-induced bias," where varying gradient magnitudes exaggerate the influence of some samples. Diffusion-ReTrac (Xie et al., 2024b) mitigates this by normalizing influence contributions. Additionally, methods not originally designed for data attribution, such as CLAP4CLIP (Jha et al., 2024) for VLMs, can still provide valuable insights through components like memory consolidation, weight initialization, and task-specific adapters that highlight crucial data points during training.

Other Miscellaneous Methods By contrasting similar and dissimilar data, these techniques trace how training examples influence model outputs. For example, one approach fine-tunes a pre-trained text-to-image model using exemplar pairs and employs NT-Xent loss to generate soft influence scores (Wang et al., 2023c). Similarly, Data Adaptive Traceback (DAT) (Peng et al., 2024) aligns pretraining examples with downstream performance in a shared embedding space. Moreover, adversarial attack studies (Wang et al., 2024c) demonstrate that intra-modal contrastive learning can be used to distinguish between adversarial and benign samples, while cross-modal loss highlights features critical for image-text alignment.

F.5 Additional Details on Feature Visualizations

Visualizing Relevance Scores For a given prediction, Robnik-Šikonja and Kononenko (2008) visualizes a relevance score of each feature by examining how the prediction changes if the feature is excluded, calculated as the probability difference before and after excluding the feature. Zintgraf et al. (2017) enhances this model by considering spatial dependence, proposing that a pixel's impact is strongly influenced by its neighboring pixels, thus expanding from pixel-level to patch-level relevance and measuring feature influences from hidden layers. Chefer et al. (2021) further improves the method of accumulating relevance across multiple layers by introducing a relevance propagation rule. Another line of work involves training a separate explanation model to predict feature relevance scores and then visualize them. Ribeiro et al. (2016) train an explanation model to evaluate the contribution of each image patch or word to the prediction. Park et al. (2018) collect two new datasets to train a multimodal model that can jointly generate visual attention masks to localize salient regions and region-grounded text rationales. Lyu et al. (2022) extends the work of (Ribeiro et al., 2016) by developing a more detailed analysis framework. They decompose a multimodal model into unimodal contributions (UC) and multimodal interactions (MI), and then apply (Ribeiro et al., 2016) method to learn relevance scores for each feature based on

Paper	Interpretability Aspect	Analyzed Component	Application
(Gandelsman et al., 2024a)	Knowledge Localization	Self-attention	Spurious Corr, Segmentation
(Balasubramanian et al., 2024)	Knowledge Localization	Self-attention	Spurious Corr, Segmentation
(Bhalla et al., 2024)	Knowledge Localization	Layers	Spurious Corr, Model Editing
(Parekh et al., 2024)	Knowledge Localization	Self-attention	-

Table 6: Additional Details on Text-Explanations of Internal Embeddings Papers

these unimodal contributions and multimodal interactions. Liang et al. (2022) further extends to be a four-stage interpretation framework: unimodal importance, cross-modal interactions, multimodal representations, and multimodal prediction.

F.6 Summary

In this section, we explore methods designed specifically to analyze the inner workings of multimodal models. Our findings reveal that the internal embeddings and neurons of models like CLIP can be interpreted using human-understandable concepts. Additionally, the cross-attention layers in text-to-image diffusion models provide valuable insights into image composition. For training data attribution and feature visualization, we observe that existing techniques for vision models have been effectively adapted for multimodal models. In Table 10, we provide a comprehensive listing and analysis of all the papers discussed in this section.

G More Insights from In-Context Learning

Recent advances in understanding the internal mechanisms of in-context learning (ICL) have revealed fascinating insights into how both language models and multi-modal models process and leverage contextual information. The interpretability methods can be categorized into five main approaches: induction heads, Markov sampling, task vectors, information flow analysis, and experimental studies.

The investigation of induction heads has primarily focused on language models, with Elhage et al. (2021) establishing a mathematical framework for transformer circuits that demonstrated how one-layer attention-only transformers can perform primitive ICL through pattern assessment. Olsson et al. (2022) further expands this understanding by analyzing induction heads in full transformer architectures, revealing a phase change early in training across various model sizes. However, there remains a notable gap in understanding how induction heads operate in multi-modal contexts, with few studies examining their role in processing visual and textual information simultaneously. In the domain of statistical learning, Edelman et al. (2024) introduced Markov Chain sequence modeling to demonstrate how transformers develop statistical induction heads that approach Bayes-optimal performance. This work, while foundational, has primarily focused on textual sequences, leaving open questions about how such statistical learning mechanisms might extend to multimodal scenarios.

Another line of in-context learning analysis is information flow analysis, which has provided particularly striking insights into the differences between language and multi-modal processing. Wang et al. (2023a) establishes that in language models, label words serve as anchors for information aggregation and distribution, quantified through saliency metrics. Zhou et al. (2024) utilizes this framework to generative VLMs by introducing a new multimodal saliency metric for visual-target information flow, revealing that cross-modal interactions primarily occur in deeper layers, contrasting with the earlier information aggregation observed in pure language models. Experimental analyses have complemented these mechanistic studies, though often without direct investigation of internal mechanisms (Chen et al., 2023). Baldassini et al. (2024) and Qin et al. (2024) have highlighted that multi-modal ICL appears to prioritize textual information over visual inputs, with multi-modal alignment serving as a key bottleneck. Overall, the analytical approaches employed in multi-modal ICL have not yet achieved the sophistication of those developed for pure language models. The complexity of lengthy input sequences poses significant computational constraints, hindering detailed investigation of the underlying mechanisms. Furthermore, while existing research has identified distinct impacts across different modalities, the practical applications of these findings remain largely unexplored in the current literature.

Methods	Paper	Models	Task	Datasets
Logit Lens	(Huo et al., 2024)	LLaVa-next, InstructBLIP	VQA	LingoQA, RS-VQA, PMC-VQA, DocVQA, VQAv2
	(Jiang et al., 2024c)	LLaVA-1.5-7B, Shikra, MiniGPT-4	Hallucination Detection	COCO 2014
	(Phukan et al., 2024)	Qwen2-VL-7B, InternLM-xcomposer2-vl-7b	Hallucination Detection, VQA	High-Quality Hallucination Benchmark, TextVQA-X
	(Zhao et al., 2024a)	LLaVA-v1.5 (13B/7B), InstructBLIP, mPLUG-owl	Identifying Unanswerable Questions	VizWiz, MM-SafetyBench
Linear Probing	(Cao et al., 2020)	ViLBERT, LXMERT, UNITER	Multimodal Fusion, Cross-modal Interaction	Visual Genome, Flickr30k
	(Dai et al., 2023)	OSCAR, VinVL, BLIP, OFA	Object Hallucination Detection	COCO Caption, NoCaps
	(Salin et al., 2022)	UNITER, LXMERT, ViLT	POS Tagging, Object Counting	Flickr30K, MS-COCO
	(Tao et al., 2024)	Kosmos-2, LaVIT, EmU, Qwen-VL	Visual-language Entailment	MS-COCO
	(Hendricks and Nematzadeh, 2021)	MMT, SMT	Verb Understanding	Conceptual Captions
	(Dahlgren Lindström et al., 2020)	VSE++, VSE-C, HAL	Linguistic Properties	MS-COCO
Sparse AutoEncoder	(Lim et al., 2024)	CLIP	Image Classification	ImageNet
	(Rao et al., 2024)	CLIP, ResNet-50	Concept Discovery	CC3M
Causal Tracing	(Basu et al., 2024c)	Stable Diffusion, IMAGEN	Knowledge Localization	–
	(Basu et al., 2024a)	LLaVa	VQA, Model Editing	VQA-Constraints
	(Basu et al., 2024b)	SD-XL, DeepFloyd	Knowledge Localization	–
	(Yu and Ananiadou, 2024c)	LLaVa	VQA, Hallucination Detection	COCO
	(Palit et al., 2023)	BLIP	Causal Tracing	COCO-QA
Task Vector	(Cohen et al.)	Diffusion Model, CLIP	Multi-concept Editing	–
	(Gandikota et al., 2025)	Stable Diffusion	Image Editing	Ostris Dataset, FFHQ
	(Baumann et al., 2024)	CLIP, T2I Diffusion	Image Editing	Contrastive Prompts
In-Context Learning	(Huang et al., 2024a)	Qwen-VL, Idefics2-8B	Many-shot Learning	VizWiz, OK-VQA
	(Zhou et al., 2024)	LLaVA, MiniGPT, Qwen-VL	Image-Content Reasoning	Emoset, CIFAR10
	(Qin et al., 2024)	OpenFlamingo, GPT4V	VQA, Classification	–
	(Mitra et al., 2025)	LLaVA, Qwen-VL	Classification, VQA	BLINK, NaturalBench
	(Luo et al., 2024)	LLaVA, Mantis-Fuyu	Instruction Transfer	–
	(Baldassini et al., 2024)	IDEFICS, OpenFlamingo	VQA, Captioning	COCO, VQAv2
Neuron-LevelDescription	(Huo et al., 2024)	LLaVA-NeXT, InstructBLIP	VQA	LingoQA, RS-VQA
	(Gandelsman et al., 2024c)	CLIP	Zero-shot Segmentation	–
	(Yu and Ananiadou, 2024c)	LLaVa	VQA	COCO
	(Tang et al., 2024)	LLaMA-2, BLOOM	–	–
	(Hintersdorf et al., 2024)	Stable Diffusion, DALL-E	Neuron Localization	–
	(Huang et al., 2024c)	Qwen-VL, Qwen-Audio	–	–
	(Schwettmann et al., 2023)	GPT-J with BEIT	Image Captioning	CC3M

Table 7: A comprehensive overview of interpretability methods for Section 2

H Additional Applications

H.1 Privacy

Data Leakage through Attacks on Specific Modalities Multimodal data privacy refers to the protection of privacy when handling data from multiple modalities, such as text, images, audio, and video. Since multimodal models process information from different sources, each modality may involve different types of sensitive data, making privacy protection more complex and crucial (Zhao et al., 2024b). Traditional data privacy aims to protect original data from leakage by isolating and encrypting it through restricted secure access, especially for the large foundation models (Rao et al., 2023). Therefore, technologies such as federated learning (Li et al., 2020) and differential privacy (Dwork, 2006) can still work well for general training. However, due to the tight interconnections between multimodal data, this means that a reverse attack using data from a specific modality could still lead to the leakage of data from other modalities, which has become a major challenge in multimodal data privacy. Ko et al. (2023) focuses on similar issues, where data leakage can occur through membership attacks. In this paper, we further summarize the privacy attributes of multimodal data and define it as cross-modal privacy. Caused by the asymmetry of the knowledge contained in multimodal data, if attackers steal data from certain key modalities, it may be sufficient to reconstruct all the information, ultimately leading to data leakage. Recent work has focused on multimodal information measurement techniques (Zhao et al., 2024b; Liu et al., 2024c), which enhance privacy protection by quantifying the correlations between data from different modalities. It significantly strengthens local privacy and effectively reduces the leverage risk in MMFMs.

Privacy Leakage through Cross-modal Access Direct data leakage is typically catastrophic, but

Paper	Interpretability Aspect	Analyzed Component	Application
(Kalibhat et al., 2023)	Knowledge Localization	Neurons	-
(Oikarinen and Weng, 2023)	Knowledge Localization	Embeddings	Spurious Correlation
(Hernandez et al., 2021)	Knowledge Localization	Neurons	Improving Robustness for IC
(Bai et al., 2024)	Knowledge Localization	Neurons	Improving Generalization for IC

Table 8: Additional Details on Network Dissect Papers. IC represents image classification.

Paper	Interpretability Aspect	Analyzed Component	Application
(Basu et al., 2024b)	Knowledge Localization	Cross-attention	Model Editing
(Neo et al., 2024)	Knowledge Flow	Cross-attention	Model Editing
(Hertz et al., 2022)	Knowledge Flow	Cross-attention	Image Editing
(Tang et al., 2022)	Knowledge Flow	Cross-attention	Visualization, Compositionality

Table 9: Additional Details on Cross-Attention Interpretability Papers

such cases are rare in practical scenarios. A more common challenge of privacy leakage occurs during the training process (Fang et al., 2024a). Reverse attacks on models for specific modalities can also lead to data leakage. Liu et al. (2024b) explore the risk in vision-language models and highlight the risks that reverse attacks on multi-modal aggregation can potentially lead to the recovery of image data. The same, this type of attack can also be initiated by the trainer, who may construct partially falsified training data to reverse-query the corresponding data from other modalities (Xu et al., 2024). To prevent such privacy leakage, a key technique is feature perturbation. By adding lightweight noise, it ensures that during multimodal information fusion, knowledge from cross-modal data cannot be easily mapped independently. This enhances the privacy level in the training process.

Unreliable Samples: Poisoning Attacks Poisoning attacks pose a significant threat to data reliability, targeting the training process by injecting maliciously altered data into the system. These attacks manipulate the training data to introduce vulnerabilities, potentially causing models to produce inaccurate predictions or exhibit unintended behaviors. Attackers usually craft subtle changes but significantly impact model performance. In multimodal models, apart from the traditional poisoning of tampering with the original data, altering the mapping relationships has become another critical attack vector. Liu et al. (2024d) learn the impact of asymmetric data attacks on model training is significant, as even a small amount of manipulated data can cause a severe decline in model performance. This also leads to more severe backdoor attacks, where attackers can execute the attack without the need for additional information injection (Liu et al., 2024a; Yang et al., 2024b). Aimed to these attacks, an effective solution is to generate adversarial examples for evaluation. By evaluating the symmetry of the modalities and the mapping relationships, toxic samples can be avoided from harming the network during training.

H.2 Other Relevant Applications

In this section, we highlight some of the other relevant applications using mechanistic insights for multimodal models:

Controlled Image Generation and Editing In text-to-image diffusion models, task vectors can be used to control and edit the intensity of a specific concept in an image (Baumann et al., 2024; Gandikota et al., 2025), while keeping other parts of the image unchanged. For example, given the prompt "*An image of a boy in front of a cafe*", if the size of the boy's eyes needs to be increased, a task vector corresponding to eye size is added to the model to modify the visual concept of the eyes. In the case of image editing, (Hertz et al., 2022) intervenes on the interpretable cross-attention features to incorporate text-guided image edits.

Zero-shot Segmentation and Mitigating Spurious Correlations The Representation Decomposition framework (Gandelsman et al., 2024a; Balasubramanian et al., 2024) enables mapping the contributions of different visual tokens to the final [CLS] token. This decomposed information can be ranked based on CLIP similarity to identify the most important tokens for a specific visual concept.

These selected tokens then form the segment representing the given concept. This framework when combined with Text-Explanations of Internal Components (see Sec.3.1), can also mitigate spurious correlations. For e.g., certain attention heads can be identified that encode spurious attributes (e.g., water when classifying waterbirds). By ablating the contributions of these attention heads to the final [CLS] token in the image encoder, spurious correlations in CLIP models can be partially mitigated.

I Tools and Benchmarks

There are many interpretability tools for LLMs covering attention analysis (Nanda and Bloom, 2022; Fiotto-Kaufman et al., 2024), SEA analysis (Joseph Bloom and Chanin, 2024), circuit discovering (Conmy et al., 2023a), causal tracing (Wu et al., 2024), vector control (Vogel, 2024; Andy Zou, 2023), logit lens (bel, 2023), and token importance (Lundberg and Lee, 2017). However, the tools for interpreting MMFMs cover narrow fields. Yu and Ananiadou (2024d); Stan et al. (2024) mainly focuses on the attention mechanism in generative VLMs. Aflalo et al. (2022) introduces a tool to visualize attentions and also hidden states of generative VLMs. Joseph (2023) proposes a tool for vision transformers, mainly focusing on attention maps, activation patches, and logit lenses. Besides, for diffusion models, Lages (2022) provides a visualization of the inner diffusion steps of generating an image.

A unified benchmark for interpretability is also a very important research direction. In LLMs, Huang et al. (2024b) introduces a benchmark for evaluating interpretability methods for disentangling LLMs' representations. Thurnherr and Scheurer (2024) presents a novel approach for generating interpretability test beds using LLMs which saves time for manually designing experimental test data. Nauta et al. (2023); Schwettmann et al. (2024) also provides benchmarks for interpretability in LLMs. However, there is no such benchmark for multimodal models, which is an important future research direction.

Overall, compared to the comprehensive tools and benchmarks in the LLMs field, there are less for multimodal foundation models. Providing a comprehensive, unified evaluation benchmark and tools is a future research direction.

Methods	Paper	Models	Task	Datasets
Text-Explanations of Internal Embeddings	(Gandelsman et al., 2024a)	CLIP	Image Retrieval, Segmentation	Waterbirds, CUB, Places, ImageNet-segmentation
	(Balasubramanian et al., 2024)	CLIP	Image Retrieval, Segmentation	ImageNet
	(Bhalla et al., 2024)	CLIP	Image Classification	CIFAR100, MIT States, MSCOCO, LAION, CelebA, ImageNetVal
	(Parekh et al., 2024)	DePALM (CLIP+OPT)	Image Classification	COCO
Network Dissection	(Oikarinen and Weng, 2023)	ResNet	Image Classification	CIFAR100, Broden, ImageNet
	(Kalibhat et al., 2023)	DINO	Image Classification	ImageNet, STL-10
	(Hernandez et al., 2021)	ResNet, Gan, AlexNet	Image Classification	ImageNet
	(Bai et al., 2024)	ResNet	Image Classification	ImageNet
Training Data Attribution Method	(Hu et al., 2024)	CLIP(ViT-B/16 + LoRA)	—	FGVC-Aircraft, Food101, Flowers102, Describable Textures Dataset(DTD), Cifar-10
	(Mlodozeniec et al., 2024)	DDPM	—	CIFAR-10, CIFAR-2, ArtBench
	(Park et al., 2023)	ResNet-9; ResNet-18; BERT	—	QNLI, CIFAR-10, ImageNet
	(Zheng et al., 2024b)	DDPM	—	CIFAR(32×32), CelebA(64×64), ArtBench
	(Xie et al., 2024b)	DDPM/DDIM	—	CIFAR-10 airplane subclass, MNIST zero subclass, ImageNet, CelebA, Artbench-2
	(Jha et al., 2024)	CLIP	—	CIFAR100, ImageNet100, ImageNet-R, CUB200, VTAB
	(Pruthi et al., 2020)	ResNet-56	—	CIFAR-10, MNIST
	(Qiu et al., 2022)	ResNet50, VGG16	—	ImageNet, Pascal VOC
	(Yang et al., 2024c)	BLIP2(blip2-opt-2.7b), instructBLIP(instructblip-vicuna-7b), LLaVA(LLaVA-v1.5-7b)	—	visualQA, CroPA
	(Zheng et al., 2024a)	CLIP	—	Flickr30, MS COCO
	(Chen et al., 2024a)	BLIP2-OPT(2.7B), LLaVA-V1.5(7B), MiniGPT-4(7B)	—	E-VQA, E-IC
	(Mitra et al., 2024)	InstructBLIP-13B, LLaVA-1.5-13, Sphinx, GPT-4V	—	Winoground, WHOOPS!, SEEDBench, MMBench, LLaVA-Bench
	(Fu et al., 2024)	PaliGemma-3B-Mix-448	—	DOCCI
	(Kwon et al., 2024)	RoBERTa / Llama-2-13B-chat, stable-diffusion-v1.5	—	MRPC, SST2, WNLI, QQP, Dreambooth (various transformations)
	(Wang et al., 2023c)	DINO, MoCov3, CLIP, ViT, ALADIN, SSCD	—	ImageNet-1K, BAM-FG, Artchive, MSCOCO
	(Peng et al., 2024)	CLIP	—	CIFAR10, CIFAR100, FGVC Aircraft, Oxfordpet, Stanford Cars, DTD, Food101, SUN397
	(Peng et al., 2024)	CLIP, OpenCLIP-G/14, EVA-02-CLIP-bigE-14-plus, ALBEF, TCL, BLIP, BLIP2, MiniGPT-4	—	MSCOCO, Flickr30K, SNLI-VE
	(Wang et al., 2023d)	CLIP	—	Conceptual Captions, MS-CXR, ROCO, RSICD
	(Fang et al., 2024b)	DensetNet-121	—	ITAC, iCTCF, BRCA, ROSMAP
Cross-attention Interpretability Methods	(Basu et al., 2024b)	SD-1.5, SD-XL, DeepFloyd	Model Editing	Concept-Editing Dataset
	(Neo et al., 2024)	LLaVA, LLaVA-Phi	Potential Application: Coarse Segmentation	COCO Detection Dataset
	(Hertz et al., 2022)	Stable-Diffusion	Image Editing	Custom Image Editing Dataset
	(Tang et al., 2022)	Stable-Diffusion	Visualization	Custom Dataset

Table 10: A comprehensive overview of interpretability methods for Section 3.