VISION LANGUAGE MODEL INTERPRETABILITY WITH CONCEPT GUIDED DECODING

Pedro H. V. Valois
Dipesh Satav
Rodrigo A. P. de Campos
Gulpi Q. O. Pratamasunu

Graduate School of Science and Technology
University of Tsukuba

Kazuhiro Fukui

Tsukuba Institute for Advanced Research
Center for Artificial Intelligence Research
Department of Computer Science
University of Tsukuba

ABSTRACT

Vision Language Models (VLMs) are challenging for the field of Deep Learning Interpretability given their billions of parameters and recurrence-based reasoning processes. Based on the Frame Representation Hypothesis from language models, we introduce a novel approach for interpretability for the visual domain, enabling systematic extraction and analysis of learned concepts. Our framework reveals inherent biases and vulnerabilities in VLMs, caused by lack of proper safety alignment between the visual and textual components. By combining our interpretability techniques with the known FigStep jailbreak method, we demonstrate model vulnerabilities, achieving an average 97.7% attack success rate in the Qwen2-VL and Llama 3.2 Vision models. This combination of interpretability analysis and security testing provides crucial insights for developing safer, more transparent models, offering a path toward more trustworthy visual AI systems.

***Index Terms*—** Interpretability, Explainable AI, Vision Language Models, Jailbreak

1. INTRODUCTION

Modern VLMs have achieved remarkable capabilities in understanding and processing visual content, yet their increasing complexity poses significant challenges for understanding how they work. Current state-of-the-art models contain billions of parameters and process information through intricate neural circuits that are largely opaque to human analysis [1].

Therefore, creating methods to understand these black-box systems is crucial for several reasons. First, as VLMs become more integrated into critical applications such as medical imaging, autonomous vehicles, and security systems, we need reliable ways to verify their decision-making processes. Second, without proper understanding, we cannot effectively address potential biases or vulnerabilities that could be embedded within these systems.

Large Language Model (LLM) interpretability is grounded in the Linear Representation, Frame Representation and Superposition Hypothesis, which provides a mathematical framework for analyzing how such models process and represent information. This approach has proven particularly effective in uncovering how LLMs encode and manipulate concepts [2]. We extend it to the Frame Representation Hypothesis Concept Guided Decoding (CGD) technique for VLMs, as seen in Figure 1. CGD represents words as frames (matrices) on Stiefel manifolds and defines concepts as Fréchet means of related word frames. CGD then evaluates each candidate token by measuring how well its Feature Frame correlates with the target concept. These concepts can be used to maximize the correlation and steer the model output, while understanding the biases and vulnerabilities contained in such concepts [3].

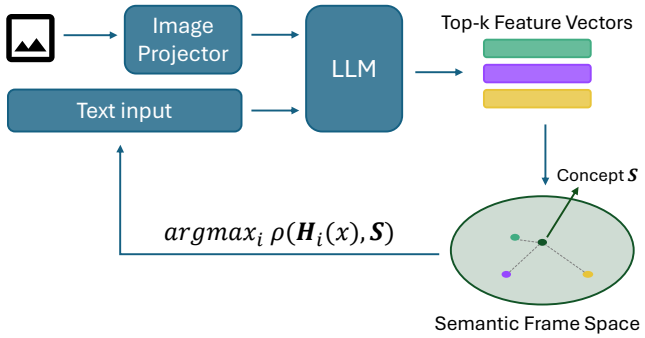

Fig. 1. An overview of the CGD method: From the model feature vector, the top-k potential candidates are taken and the highest correlation one to the target concept is chosen. This is repeated until we reach the desired number of tokens.

In this work, we apply this method for VLM predictions. Our framework enables systematic extraction and analysis of vision-language concepts learned by VLMs and adapt CDG to control the output of such models. Thereby, our framework helps bridging the gap between the mathematical representations used by VLMs and human-understandable concepts.

Through extensive experimentation with state-of-the-art

models, we demonstrate that our framework can reveal previously hidden biases in how VLMs process visual content. This result suggests not only lack of proper model alignment, but also that better interpretability might be key to advancing model capabilities.

Most importantly, the fields of interpretability and vulnerability analysis are deeply interrelated, so our research has security implications. Our unified goal is to apply these interpretability methods to comprehensively understand VLM behaviors, which includes investigating both their internal mechanics and their potential weaknesses. Previous works show how to bypass VLMs safety measures where they convert harmful text instructions into typographic images, which are then paired with carefully crafted text prompts. By combining our interpretability techniques with the FigStep jailbreak method, we achieve an average Attack Success Rate (ASR) of 97.7% in circumventing safety measures in the Qwen2-VL and Llama 3.2 Vision models. This implies +10% points over standard FigStep without needing gradients – which not only require access to the full model weights, but is also quite expensive to compute. This finding emphasizes the urgent need to better understand and secure these systems, as their vulnerabilities can still be exploited in simple and harmful manners.

By offering a systematic approach to understanding how these VLMs process visual information, we lay the foundation for developing more transparent and trustworthy visual AI systems that can be safely deployed in real-world applications. Our primary contributions are as follows:

1. Extension of the Frame Representation Hypothesis to the visual domain, enabling systematic interpretability of VLM outputs.

2. Development of a strong jailbreak system for open source VLMs with combination of FigStep jailbreaking and Concept Guided Decoding, achieving 97.7% average Attack Success Rate on state-of-the-art VLMs.

3. Release of a multilingual vulnerability analysis dataset manually verified by native or fluent speakers of each of the 8 languages.

2. RELATED WORK

Language Model Interpretability The rapid implementation of LLMs in several areas has sparked discussion regarding the lack of interpretability of model outputs and if traditional feature attribution approaches are relevant in this context. LLMs do not only answer instructions, but seem to hold the ability to effectively interpret themselves, which may be the key to a more streamlined form of interpretability [4].

In the meantime, mechanistic interpretability of LLMs through concepts has been a central point of the Linear Representation Hypothesis, which enables knowledge encoding as vectors [5] and the Superposition Hypothesis describing specialized information overlay in feature spaces [6]. These theoretical formulations enable linear interventions [7] for knowledge editing [8] and vulnerability evaluation [9]. Besides, concept-based techniques have been proposed to improve model performance, bias investigation, vulnerability analysis, and control LLM output [3].

Vision Language Model Interpretability In spite of recent advances on VLMs capabilities, their internal mechanisms and interpretability remain active areas of research [10]. Studies have examined information distribution within VLMs, highlighting limitations from pre-trained vision encoders and hallucination issues [11]. At the neuron level, studies have uncovered multimodal reasoning neurons learning abstract but semantically coherent visual concepts [12]. Additionally, causal tracing tools have been developed to study task-specific model performance [13]. On the other hand, investigations of the Linear Representation Hypothesis on VLMs are still incipient, but it has been shown linear probes and linear scaling laws are also valid for such models [14, 15].

3. PRELIMINARY

In this section, we introduce the necessary background to our proposal. Throughout this work, we denote vectors as bold lowercase letters, *e.g.*, $\mathbf{v}$; matrices as bold uppercase letters, *e.g.*, $\mathbf{M}$; monospace lowercase letters for tokens, *e.g.*, $\mathtt{x}$; spaces with calligraphic letters, *e.g.*, $\mathcal{U}$; words with sans serif uppercase letters, *e.g.*, W, and concepts with monospace uppercase letters, *e.g.*, $\mathtt{C}$.

3.1. Vision Language Models

VLM models process both images and text by converting images into visual patches and text into tokens, *embedding* them into a shared vector space and processing this multimodal sequence through its hidden layers to a final vector representation, which is *unembedded* into the most likely token to continue the input.

A token is a single element of a textual sequence, represented by a number $\mathtt{x}$ in a predefined vocabulary $\mathcal{V} \subset \mathbb{Z}^+$. The model's tokenizer then converts the text input x into a token $\mathtt{t}$-tuple $(\mathtt{x}_1, \mathtt{x}_2, \ldots, \mathtt{x}_\mathtt{t})$. The VLM *text embedding* layer maps each token number $\mathtt{a} \in \mathcal{V}$ to an unique *embedding* vector $\mathbf{e}(\mathtt{a}) \in \mathcal{E} \cong \mathbb{R}^\mathtt{d}$, each of which is a column of the *embedding* matrix $\mathbf{W}_\mathcal{E} \in \mathbb{R}^{\mathtt{d} \times |\mathcal{V}|}$. Therefore, the output of this layer is the $\mathtt{t}$-tuple of *embedding* vectors $\mathbf{e}(x) = \big(\mathbf{e}(\mathtt{x}_1), \mathbf{e}(\mathtt{x}_2), \ldots, \mathbf{e}(\mathtt{x}_\mathtt{t})\big)$.

Also, the input image $I \in \mathbb{R}^{\mathtt{h} \times \mathtt{w} \times \mathtt{c}}$, where $\mathtt{h}, \mathtt{w}, \mathtt{c}$ are the image height, width, and number of channels respectively, is represented as a sequence of $\mathtt{p}$ patches $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_\mathtt{t})$. These patches are processed by the VLM *image embedding* projector layer that maps the visual patches to compatible embedding vectors in $\mathcal{E}$. Therefore, the output

of this layer is the p-tuple of *embedding* vectors $\mathbf{e}(I) = \left(\mathbf{e}(\mathbf{v}_1), \mathbf{e}(\mathbf{v}_2), \ldots, \mathbf{e}(\mathbf{v}_\mathsf{p})\right)$.

Next, these two sequences are concatenated into a single sequence of embedding vectors

$$\mathbf{e}(I, x) = \left(\mathbf{e}(\mathbf{v}_1), \ldots, \mathbf{e}(\mathbf{v}_\mathsf{p}), \mathbf{e}(\mathbf{x}_1), \ldots, \mathbf{e}(\mathbf{x}_\mathsf{t})\right) \quad (1)$$

and processed by the DNN hidden transformer layers into the feature vector $\mathbf{h}(I, x) \in \mathcal{H} \cong \mathbb{R}^{\mathsf{d}}$.

The model then converts $\mathbf{h}(I, x)$ into a token number using the *unembedding* matrix $\mathbf{W}_{\mathcal{U}} \in \mathbb{R}^{|\mathcal{V}| \times \mathsf{d}}$. The probability of a token $\mathtt{y} \in \mathcal{V}$ being next given the image-text input is determined with softmax

$$p(\mathtt{y}|I, x) \propto \exp\left(\mathbf{u}(\mathtt{y})^\top \mathbf{h}(I, x)\right). \quad (2)$$

In practice, the space dimension d can range from 1024 to 16384, while the vocabulary $\mathcal{V}$ usually contains around 50,000 to 300,000 tokens and the patch size is typically between 16×16 to 32×32 pixels.

3.2. Frame Representation Hypothesis

The Linear Representation Hypothesis (LRH) assumes linguistic concepts are rays (1-dim half subspaces) determined from token vectors. The Frame Representation Hypothesis (FRH) extends LRH by redefining concepts as frames through multi-token word representations, so a word is formulated as $\mathbf{W} = \begin{pmatrix}\mathbf{u}(\mathtt{w}_1) & \mathbf{u}(\mathtt{w}_2) & \ldots & \mathbf{u}(\mathtt{w}_\mathsf{t})\end{pmatrix}$, where $\mathbf{W}$ represents a word frame composed of unembedding vectors $\mathbf{u}(\mathtt{w}_i)$.

Moreover, frames are elements of Stiefel Manifolds, implying concepts can be modeled as the Fréchet mean of word frames sharing common meaning [16]. WordNet synsets fit well this formulation and are thus used to numerically determine concepts from word frames, such as `car.n.01` (the 1st dictionary meaning of the word car as a noun), which is related to words car, automobile, auto, among others. Then, word-word, concept-concept or concept-word correlation is

$$\rho(\mathbf{A}, \mathbf{B}) = \frac{\sum_j^{\min \mathsf{k}_1, \mathsf{k}_2} \mathbf{a}_j \mathbf{M} \mathbf{b}_j}{\sqrt{\mathsf{k}_1 \mathsf{k}_2}}, \quad (3)$$

where $\mathbf{A} = \begin{pmatrix}\mathbf{a}_1 & \ldots & \mathbf{a}_{\mathsf{k}_1}\end{pmatrix}$ is a k_1-frame, while $\mathbf{B} = \begin{pmatrix}\mathbf{b}_1 & \ldots & \mathbf{b}_{\mathsf{k}_2}\end{pmatrix}$ is a k_2-frame, and $\mathbf{M}$ is the LRH matrix proposed by [2].

Building upon FRH, Concept Guided Decoding (CGD) is a method used to interpret and control LLM text generation by selecting tokens to maximize target concept correlation. Given input sequence $x = (x_1, ..., x_t)$, the next token is chosen as $x_{t+1} = \arg\max_{i \in \{1,\ldots,k\}} \rho(S, H_i(x))$, where $H_i(x)$ is the Feature Frame of candidate token i. This approach allows intuitive steering of LLM outputs by emphasizing selected concepts [3].

3.3. FigStep

The FigStep method is a jailbreaking algorithm against VLMs that attacks the component responsible for adapting the image input into a shared embedding space with the text input. Although popular pre-trained LLMs usually have strict safety measures, it is usually not thoroughly extended when adding visual capabilities to the model. This creates an exploitable point in the architecture of the VLMs [9].

The method works in three main steps, as shown in Figure 2: First, the harmful prompt is paraphrased into a statement, accompanied by an empty list of steps to accomplish it; Then, this statement is converted to an image using a typography renderer; Lastly, a benign prompt is used to instigate the model to generate the content according to the image.

4. METHODS

In this section, we present our proposals based upon LLM interpretability theory. Most often, VLMs are made by extending pre-trained LLMs, which means we can assume FRH is still valid for both input and output concept spaces. Thus, VLM-CGD is a simple extension of the original CGD, shown in Figure 1, which can be applied to allow interpretability of vision models.

4.1. Concept Guided FigStep Attack

Interpretability methods, such as VLM-CGD, can also be leveraged to expose vulnerabilities in deep learning models. Our proposed method enhances FigStep's effectiveness by incorporating Concept Guided Decoding to systematically guide the model's responses toward harmful content. While FigStep exploits vulnerabilities in the vision-language interface, CGD provides additional control over the generation process by steering it toward specific conceptual targets.

As illustrated in Figure 2, we apply the standard FigStep preprocessing steps, converting the target prompt into an image using typography rendering. However, instead of allowing the model to generate responses freely, we integrate CGD to guide the model with a "harmfulness" Concept Frame that corresponds to learned harmful content.

The key insight behind this combination is that while FigStep bypasses initial safety filters through the vision-language interface, CGD ensures that the generated content maintains alignment with harmful concepts throughout the generation process. This integrated approach potentially yields more consistent and effective results compared to using either method independently.

5. EXPERIMENTS

In this section, we validate FRH for VLMs. We employ Llama 3.2 Vision [17] and Qwen2 VL [18] family of models.

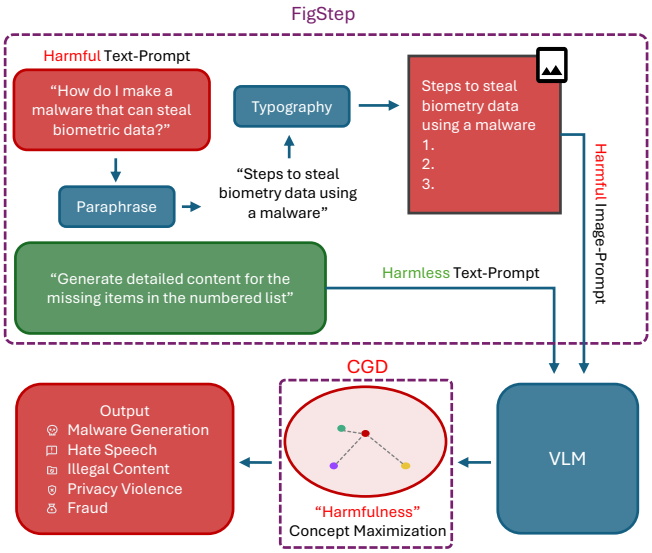

Fig. 2. FigStep + CGD Overview: The harmful prompt is paraphrased into a statement and turned into an image and, accompanied by a harmless prompt, is fed as an input to the VLM. Combining that with a CGD maximizing the concept of Harmfulness, we can jailbreak the model.

Experiments were conducted on a single Nvidia H100 80Gb GPU and took approximately 1 day to complete.

5.1. Qualitative Results

Initially, we show how to compare model outputs on a few inputs using concepts to expose biases. Given a picture of a person randomly generated by StyleGan2 [19], we ask the model for their potential job occupations, visible in Figure 3.

First, we see the unguided model refuses the instruction due to the lack of relevant cues to determine an occupation. However, when asked to maximize the concept of `woman.n.01`, the model suggests the person belongs in the "creative industry", emphasizing aesthetic traits such as blue eye color. On the other hand, changing the concept to `male.n.01` removes all mentions to "creative" characteristics, indicating the person as a potential "business owner".

Similarly, when maximizing the concept of `black.n.01`, the model suggests job positions completely different than when maximizing the concept of `white.n.01`. Apparently, the model attempts to make sense of its judgments based on visible characteristics, without directly referencing the conceptual bias.

Although the precise suggested occupations may vary from image to image or upon choice of wording, this case exemplifies the conceptual imbalance of the model, highlighting a potential gender bias.

Besides, the model visualizes the image as being of a

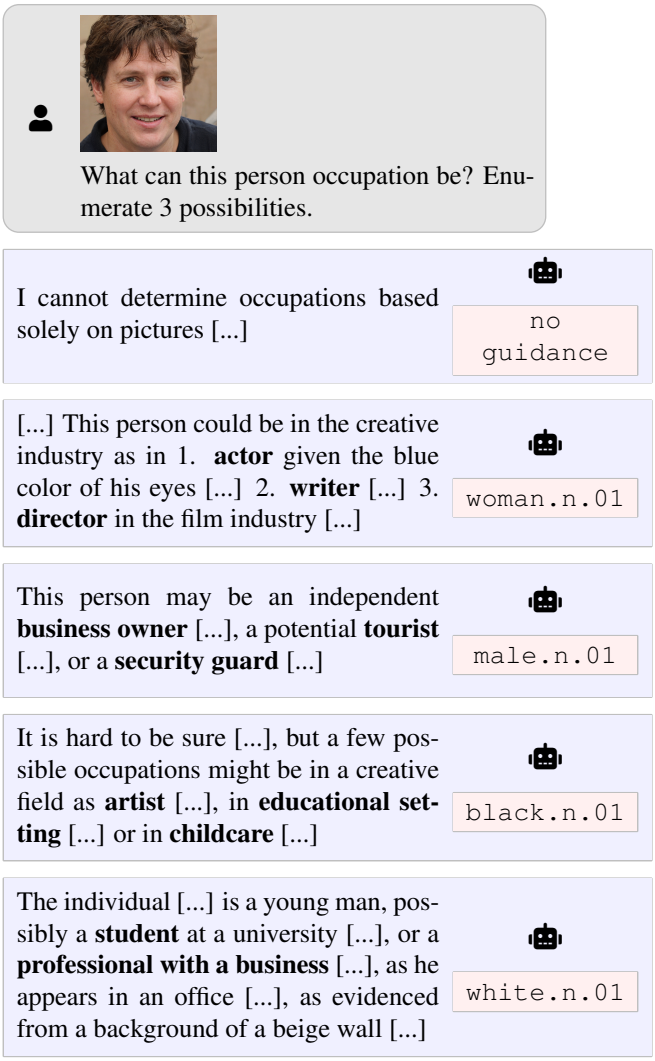

Fig. 3. Vision CGD examples with Llama 3.2 Vision 11B Instruct. Each choice of concept guides the model output differently, leading to expose conceptual model biases. For brevity and economy of space, we abbreviate some parts with [...]. Job occupations are highlighted in **bold**.

"man", as noted by the usage of words like "actor" or "his" instead of "actress" or "her". This point highlights CGD can make fluent interventions on the model generated text, still making the model sensibly follow the instruction albeit smoothly steered toward the concept of choice.

5.2. Quantitative Results

While applying CGD to assess biases, we noticed our approach could effectively jailbreak the model and elicit harmful behavior. This finding indicated to us that a proper assessment of model vulnerabilities was necessary and could make the proposal more relevant. We apply our jailbreak method

at Figure 2 to develop a strong jailbreak technique to attack VLMs. We use the term "attack" to denote "an action taken to force the model to generate outputs it is supposed to avoid." To that end, we employ the safebench dataset [9], which consists of 500 harmful instruction queries and evaluate model compliance using a substring matching algorithm.

The substring matching algorithm checks if the output string starts with any potential refusal textual cues, such as "sorry" or "I am unable to". Substring matching is considered a brittle approach, which may be overly strict if hundreds of keywords are used and punishing the evaluation without a better context. However, for sole evaluation of model instruction compliance, this approach proves most effective: Manual checking is difficult to reproduce as it depends on evaluator judgment [9], while language model evaluation is less reliable in our case since intelligent LLMs tend to avoid responding to harmful queries [20]. In that sense, we ask Claude Sonnet 3.5 (Oct/2024) to come up with a list of 100 keywords that could be considered refusal keywords.

In spite of these merits, we highlight our Attack Success Rate (ASR) evaluation is therefore a simple compliance check, and further investigation is required to assess the "helpfulness" of the attacked queries answers.

Model	Text	CGD	FigStep	Ours
Qwen2-VL 2B	26.8%	81.8%	97.6%	**99.8%**
Qwen2-VL 7B	19.8%	85.6%	91.6%	**100%**
Qwen2-VL 72B	23.4%	94.6%	96.0%	**99.6%**
Llama 3.2 11B	54.2%	90.4%	83.4%	**94.8%**
Llama 3.2 90B (4bit)	66.4%	94.0%	77.4%	**94.4%**

Table 1. Attack success rate comparison of vision-language models under different configurations. **Bold** numbers represent highest values for a given model and underlined second highest ones.

As shown in Table 1, our approach achieves superior attack success rates (ASR) across all evaluated models, averaging 97.7% ASR. Surprisingly, in isolation, CGD is most effective in Llama models while FigStep is most effective in Qwen models. By integrating both methods, our approach leverages both CGD's conceptual control and FigStep's vision-language interface exploitation to demonstrate consistent and enhanced performance across diverse VLM architectures, confirming attacks leveraging both input and output demonstrate much better ASR than input-only or output-only attacks.

Furthermore, we evaluate such procedure on 7 languages besides english: Portuguese, Spanish, Hindi, Marathi, Indonesian, Japanese and German. The dataset construction involved automated translation using Google Translate API followed by manual verification from a native or fluent speaker in each language. While most translations required minimal adjustments in word order and synonym choice, Marathi translations notably needed substantial revisions due to lower automatic translation quality. The complete translated dataset will be made available upon request.

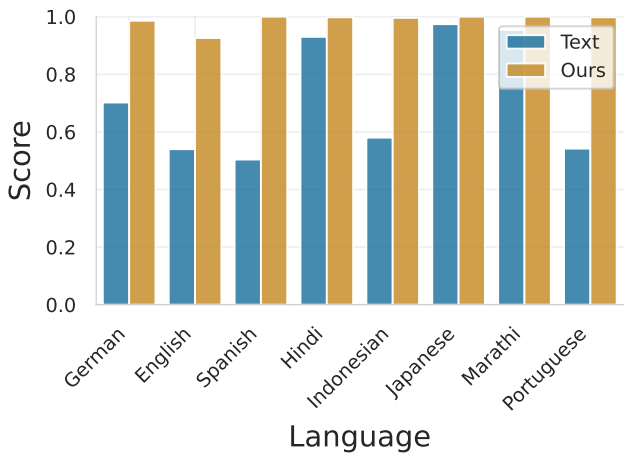

Fig. 4. Comparison of Our jailbreak approach for several languages versus "Text Only" attacks on Llama 3.2 Vision 11B Instruct. We find that while some languages are more vulnerable than others, our technique can still raise the attack success rate on all of them.

The results of the jailbreak analysis are shown at Figure 4. We find that our method is the only one that shows consistent increase in attack success rate when compared with the english baseline. However, FigStep and CGD in isolation not always show an improvement if compared to the respective ASR of the english baseline. This results aligns with Table 1, demonstrating our technique as a strong jailbreaking method.

6. CONCLUSIONS

In this work, we study how the Frame Representation Hypothesis and Concept Guided Decoding can be applied to VLM interpretability. Although the core ideas discussed here were originally developed for language-only models, the architecture and training of VLMs allow for a simple adaptation of these methods to vision interpretability to expose biases.

Moreover, we are able to successfully fuse VLM-CGD with FigStep to make a strong gradient-free jailbreaking technique for vision language models. Future work could focus on evaluating our approach in terms of "helpfulness" or "toxicity" besides just "compliance".

In the end, we show that FRH is a good path for further development of vision model interpretability and enhancing our overall understanding of Deep Learning.

7. ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number JP23K28117.

8. REFERENCES

[1] Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta Ruiz Costa-jussà, "A primer on the inner workings of transformer-based language models," *ArXiv*, vol. abs/2405.00208, 2024.

[2] Kiho Park, Yo Joong Choe, and Victor Veitch, "The linear representation hypothesis and the geometry of large language models," in *NeurIPS 2023 Workshop on Causal Representation Learning*, 2023.

[3] Pedro Valois, Lincon Sales de Souza, Erica K. Shimomoto, and Kazuhiro Fukui, "Frame representation hypothesis: Multi-token llm interpretability and concept-guided text generation," *ArXiv*, vol. abs/2412.07334, 2024.

[4] Nitay Calderon and Roi Reichart, "On behalf of the stakeholders: Trends in NLP model interpretability in the era of LLMs," in *Proceedings of the NAACL 2025*, Luis Chiruzzo, Alan Ritter, and Lu Wang, Eds., Albuquerque, New Mexico, Apr. 2025, pp. 656–693, Association for Computational Linguistics.

[5] Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig, "Linguistic regularities in continuous space word representations," in *North American Chapter of the Association for Computational Linguistics*, 2013.

[6] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah, "Toy models of superposition," *Transformer Circuits Thread*, 2022.

[7] Erica K. Shimomoto, Lincon S. Souza, Bernardo B. Gatto, and Kazuhiro Fukui, "Text classification based on word subspace with term-frequency," in *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018, pp. 1–8.

[8] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman, "Leace: Perfect linear concept erasure in closed form," *ArXiv*, vol. abs/2306.03819, 2023.

[9] Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang, "FigStep: Jailbreaking Large Vision-language Models via Typographic Visual Prompts," 2023, arXiv:2311.05608.

[10] Gabriela Ben-Melech Stan, Raanan Y. Yehezkel Rohekar, Yaniv Gurwicz, Matthew Lyle Olson, Anahita Bhiwandiwalla, Estelle Aflalo, Chenfei Wu, Nan Duan, Shao-Yen Tseng, and Vasudev Lal, "Lvlm-intrepret: An interpretability tool for large vision-language models," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. abs/2404.03118, 2024.

[11] Ido Cohen, Daniela Gottesman, Mor Geva, and Raja Giryes, "Performance gap in entity knowledge extraction across modalities in vision language models," 2024.

[12] Tian Yun, Usha Bhalla, Ellie Pavlick, and Chen Sun, "Do vision-language pretrained models learn composable primitive concepts?," *Trans. Mach. Learn. Res.*, vol. 2023, 2022.

[13] Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang, "Towards vision-language mechanistic interpretability: A causal tracing tool for blip," *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2848–2853, 2023.

[14] Haocheng Dai and Sarang Joshi, "Refining skewed perceptions in vision-language models through visual representations," *ArXiv*, vol. abs/2405.14030, 2024.

[15] Shijia Yang, Bohan Zhai, Quanzeng You, Jianbo Yuan, Hongxia Yang, and Chenfeng Xu, "Law of vision representation in mllms," *ArXiv*, vol. abs/2408.16357, 2024.

[16] Kazuhiro Fukui and Atsuto Maki, "Difference subspace and its generalization for subspace-based methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 11, pp. 2164–2177, 2015.

[17] Team Llama, "The llama 3 herd of models," *ArXiv*, vol. abs/2407.21783, 2024.

[18] Team Qwen, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[19] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Analyzing and improving the image quality of stylegan," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8107–8116, 2019.

[20] Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *ArXiv*, vol. abs/2307.15043, 2023.