

What Do VLMs NOTICE? A Mechanistic Interpretability Pipeline for Gaussian-Noise-free Text-Image Corruption and Evaluation

Michal Golovanesky^{*1}, William Rudman^{*1},
Vedant Palit², Ritambhara Singh¹, Carsten Eickhoff⁴
¹ Brown University, ² IIT Kharagpur, ³ University of Tübingen
{michal_golovanevsky, william_rudman}@brown.edu

Abstract

Vision-Language Models (VLMs) have gained prominence due to their success in solving complex cross-modal tasks. However, the internal mechanisms of VLMs, particularly the roles of cross-attention and self-attention in multimodal integration, are not fully understood. To address this gap, we introduce NOTICE, a Gaussian-Noise-free Text-Image Corruption and Evaluation pipeline for mechanistic interpretability in VLMs. NOTICE introduces Semantic Image Pairs (SIP) corruption, the first visual counterpart to Symmetric Token Replacement (STR) for text. Through NOTICE, we uncover a set of “universal attention heads” in BLIP and LLaVA that consistently contribute across different tasks and modalities. In BLIP, cross-attention heads implement object detection, object suppression, and outlier suppression, whereas important self-attention heads in LLaVA only perform outlier suppression. Notably, our findings reveal that cross-attention heads perform image-grounding, while self-attention in LLaVA heads do not, highlighting key differences in how VLM architectures handle multimodal learning*.

1 Introduction

Vision-Language Models (VLMs) have become central to both computer vision and natural language processing (NLP) due to their ability to complete complex multimodal tasks like visual recognition (Han et al., 2023; Menon and Vondrick, 2022), visual question answering (Liu et al., 2024; Bai et al., 2023; Li et al., 2022, 2020), and image captioning (Lu et al., 2019; Yang et al., 2023; Zhang et al., 2021). Many foundational VLMs, such as BLIP (Bootstrapping Language-Image Pre-training) (Li et al., 2022), use cross-attention to

capture interactions between modalities (Tan and Bansal, 2019; Lu et al., 2019; Li et al., 2022), while others, like LLaVA (Liu et al., 2024), successfully use self-attention for multimodal integration. Despite recent works investigating the differences in cross-attention and self-attention in VLMs (Ilinykh and Dobnik, 2022a; Parcalabescu et al., 2020; Djamila-Romaissa et al., 2022), the specific role of different attention heads in vision-language is not fully understood, particularly in large VLMs such as LLaVA (Yuksekgonul et al., 2022).

In NLP, mechanistic interpretability (MI) has proven effective for unraveling the complex structure of LLMs by identifying pathways, or *circuits*, that enable behaviors like logical reasoning and indirect object identification (Hanna et al., 2023; Wang et al., 2022). A core method in MI is *causal mediation analysis* (CMA), which measures output changes by intervening on inputs and tracking effects. To apply CMA in Transformers, Meng et al. (2023) use *activation patching*, which identifies key model components for a given task by replacing the hidden states of a *corrupt* run with those from a *clean* run. A “clean run” refers to a standard forward pass on an input. A “corrupt” run either alters the input by swapping input tokens with semantically related tokens or applies Gaussian noise to token embeddings so that the model can no longer produce the correct answer from the clean pass. While successful in NLP, mechanistic interpretability methods are understudied in VLMs due to the need for an effective semantic-based corruption scheme for images.

Previous works applying activation patching to VLMs only study BLIP, focus on Multi-Layer Perceptron (MLP) layers in BLIP’s decoder, and apply Gaussian noise to image embeddings to create corrupt runs (Palit et al., 2023). However, multimodal interaction occurs in BLIP’s image-grounded text

*Authorship order determined by coin flip.

*Code: <https://github.com/wrudman/NOTICE/>

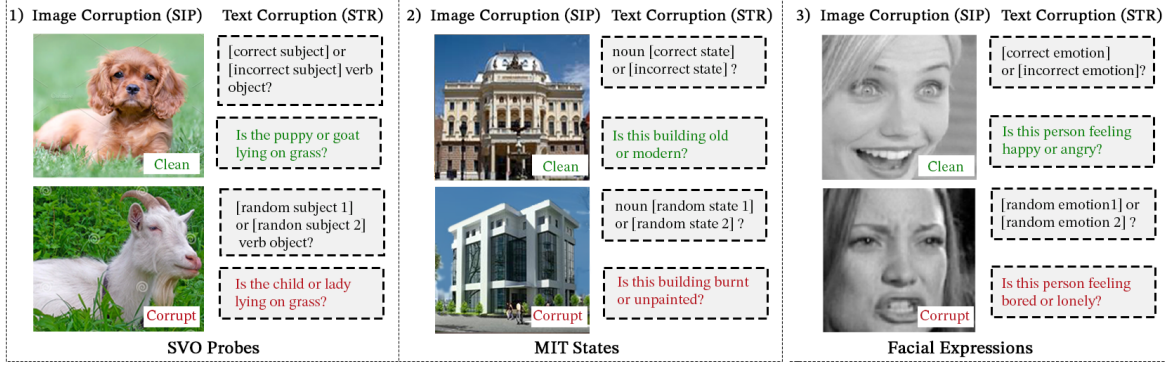


Figure 1: NOTICE applied to SVO-Probes, MIT-States, and Facial Expression Recognition. NOTICE involves creating Semantic Image Pairs for image corruption and Symmetric Token Replacement for text corruption.

encoder through the cross-attention layers, meaning findings from Palit et al. (2023) do not provide insights on how VLMs integrate modalities. Further, recent works have shown that creating corrupt runs by applying Gaussian noise to token embeddings can produce illusory results (Zhang and Nanda, 2024). Zhang and Nanda (2024) advocate for *symmetric token replacement* (STR) that swaps tokens in a clean prompt with semantically related tokens to create a corrupt run as it keeps the corrupted run within distribution and provides more reliable insights. Current approaches to image corruption lack an equivalent method to STR that does not use Gaussian noise for image corruption, leaving a gap in reliable multimodal analysis.

We propose *Semantic Image Pairs* (SIP) corruption, which leverages existing datasets to create clean and corrupt images that vary by one semantic property (Figure 1). For instance, in the SVO-Probes dataset (Hendricks and Nematzadeh, 2021), subject-verb-object triplets are already curated to vary by just one element (e.g., ‘puppy, lying, grass’ vs. ‘goat, lying, grass’), allowing us to create “clean” and “corrupt” image pairs from the existing image labels (Figure 1 panel 1). SIP is not limited to curated datasets and can be extended to any image dataset with easily interchangeable attributes, such as cars, shapes, animals, etc. Further, we show in Section 4.1 that SIP corruption can be performed using generative models to create corrupt images.

Our framework, NOTICE (Gaussian-**NO**ise-free **T**ext-**I**mage **C**orruption and **E**valuation), is a mechanistic interpretability pipeline for VLMs. NOTICE uses our novel SIP method to corrupt images and STR to corrupt text. Using NOTICE, we uncover “universal cross-attention heads” in BLIP and “universal self-attention heads” in LLaVA that

consistently produce large patching effects across all tasks. We show that universal cross-attention heads implement human-interpretable functions categorized into three classes: object detection, object suppression, and outlier suppression. We find that both self-attention and cross-attention heads implement outlier suppression. However, only cross-attention heads in BLIP perform the image-grounding functions of object detection and object suppression, highlighting key differences in how VLM architectures handle multimodal learning.

2 Related Work

Multimodal Interpretability. Despite the successes of VLMs, the role of attention in multimodal integration remains under-studied (Xu et al., 2023). Most works either design probing tasks to identify what vision-language concepts models struggle to capture (Shekhar et al., 2017a; Hendricks and Nematzadeh, 2021; Lu et al., 2023) or assess whether models rely more on one modality over the other (Cao et al., 2020; Salin et al., 2022; Hessel and Lee, 2020). Probing tasks evaluate whether a model encodes a concept by training a linear classifier on its hidden states. Salin et al. (2022) create visual, text, and multimodal probing tasks to show that models heavily rely on text priors and struggle with object, size, and counting. Gradient-based methods analyze gradients to understand how changes in input features influence outputs, assigning importance scores to features and providing insights into the decision-making process. Liang et al. (2023) use second-order gradients, building on LIME (Ribeiro et al., 2016) and Shapley values (Merrick and Taly, 2020), to disentangle cross-modal interactions and determine how much each modality contributes to decisions. While these works have provided

insights into multimodal models, recent research shows that 1) probing methods can produce misleading results (Belinkov, 2022) and 2) gradient-based methods for feature attributions can fail to predict behavior (Bilodeau et al., 2024). Instead, Giulianelli et al. (2021); Conneau et al. (2018) demonstrate that causal approaches are more informative than probing representations.

Multimodal models typically integrate their inputs through one of two main approaches: (1) early fusion (Liu et al., 2024; Li et al., 2019), where self-attention processes combined image patches and text representations, or (2) cross-attention fusion (Tan and Bansal, 2019; Lu et al., 2019), where queries from one modality interact with representations from the other. (Frank et al., 2021) investigate cross-modal integration using input ablation techniques and conclude that self-attention models, such as VisualBERT, show less sensitivity to missing visual input compared to cross-attention models like ViLBERT, which exhibit more significant reliance on cross-modal interactions. (Ilinykh and Dobnik, 2022b) finds that vision-text cross-attention learns visual grounding of noun phrases into objects, while text-to-text attention captures low-level syntactic word relations. (Parcalabescu et al., 2020) probes dual-stream VLMs through tasks like image-sentence alignment and counting, revealing strong alignment performance but limited grounding of visual symbols, resulting in poor counting capabilities. However, there has yet to be a comprehensive study applying mechanistic interpretability techniques to investigate how recent architectures like BLIP and LLaVA 1) use attention mechanisms to integrate multimodal information and 2) determine the specific functions that attention heads in these models learn to perform.

Causal Mediation Analysis and Mechanistic Interpretability. Causal mediation analysis (CMA) uncovers how interventions in one part of a system cause changes in another. Recent works have adapted CMA to study language models by using *activation patching* (Meng et al., 2023) to identify causal pathways, or *circuits*, responsible for certain behaviors in the model. Activation patching shows the indirect effect of a component on a model’s output by corrupting key input tokens and then restoring outputs by patching the internal modules from the clean run to the corrupted run. This approach has revealed that attention heads in transformers implement human-interpretable algorithms for “greater than” tasks (Hanna et al., 2023), indi-

rect object identification (Wang et al., 2022), and learning mathematical group operations (Chughtai et al., 2023). Evidence suggests that models implement similar circuits across various tasks, indicating that studying a small set of computational units may provide insights into model behavior (Gould et al., 2023; Merullo et al., 2024). Although works in mechanistic interpretability have provided critical insights into LLMs, works applying activation patching to multimodal models have been limited due to the absence of an effective semantic-based method for image corruption. Given that Zhang and Nanda (2024) demonstrate Gaussian noise corruption can lead to “illusory” patching results, a semantic-based corruption scheme for images is needed to apply causal mediation analysis to VLMs. In this work, we introduce a semantic-based image corruption method (SIP) and token replacement (STR) for text, enabling meaningful causal mediation analysis to reveal distinct roles of self-attention and cross-attention in multimodal models like BLIP and LLaVA, advancing our understanding of vision-language integration.

3 Methods

NOTICE: Overview The first step in activation patching is developing a method to corrupt the input, preventing the model from predicting the correct answer without causing representations to be out-of-distribution. We propose the first semantic-based dual corruption scheme for VLMs to enhance understanding of each modality’s contribution to VQA tasks. We assess the impact of corrupting input text with symmetric token replacement (Wang et al., 2022) and images using our novel *Semantic Image Pairs* (SIP) framework.

Model Selection For our model selection, we use LLaVA-1.5-7B (Vicuna base) (Liu et al., 2024) and BLIP-VQA-Base (Li et al., 2022), chosen for their relevance and alignment with recent interpretability research. BLIP-VQA was included to enable direct comparison with prior work applying activation patching to vision-language models (Palit et al., 2023), facilitating a meaningful evaluation of our corruption scheme (NOTICE) against Gaussian-noise patching. In addition, we also focus on LLaVA-1.5 to align with more recent interpretability studies, including (Luo et al., 2024), (Neo et al., 2024), and (Jiang et al., 2024).

Text Corruption: Symmetric Token Replacement. Corrupting text inputs with symmetric token

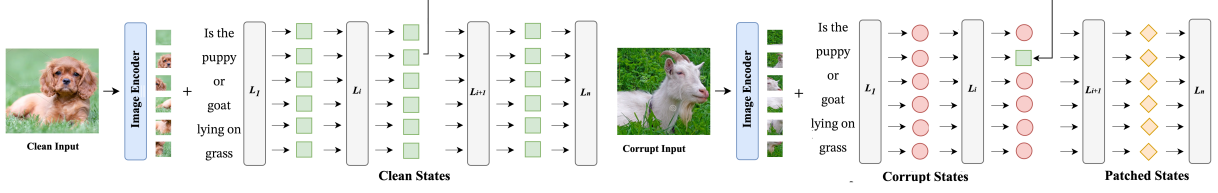


Figure 2: Activation Patching using SIP corruption. The image of the puppy is the clean image, I , and the goat is the corrupt image, I^* . Patching the correct answer token “puppy” at $M_I(I, T)$, from the clean image to the “puppy” token at $M(I^*, T)$ creates the patched states $M'(I^*, T)$ shown as orange diamonds.

replacement (STR) swaps out the correct answer token of a prompt with another similar token that has the same length (Wang et al., 2022). For indirect object identification, Wang et al. (2022) swap the names of the indirect and direct objects with randomly sampled names. The goal of STR is to corrupt inputs so the language model cannot complete the task correctly while avoiding Gaussian noise corruption that pushes token embeddings off-distribution. For our VQA tasks, we swap both multiple-choice options in the input prompt with two randomly sampled incorrect multiple-choice options with the same tokenized length from a different sample in the dataset. Figure 1 illustrates how STR alters the input text for a prompt.

Image Corruption: Semantic Image Pairs.

While STR is a widely used text corruption method, extending semantic corruption to the vision domain presents significant challenges as it requires a semantically aligned image pair. To address this, we propose SIP corruption, a visual corollary to STR, which modifies only one semantic concept in an image (see Figure 1). We implement SIP in three diverse image-classification datasets: SVO-Probes (Hendricks and Nematzadeh, 2021) (an extension to FOIL it! (Shekhar et al., 2017b)), MIT States (Isola et al., 2015), and Facial Expressions (Goodfellow et al., 2013). While natural variation in backgrounds can happen between semantic image pairs, we select datasets where nouns, subjects, verbs, and objects are known and can be controlled. The Facial Expressions dataset was curated to include only black-and-white photos of faces expressing different emotions. In MIT-States, object-state pairs are known and can be used to maintain objects (e.g., “building”) while varying their states (e.g., “modern”). The SVO-Probes dataset is designed to vary only one element (subject, verb, or object) at a time (e.g., only varying “puppy” and “goat”, while maintaining “grass” and “lying” constant, as seen in Figure 1 panel 1). Thus, for each dataset, we col-

lect the known elements (emotions, objects/states, subjects/verbs/objects) and create image pairs that differ on that element. Our framework can be extended to any image dataset with clearly defined attributes, such as cars, shapes, animals, etc., using the class labels to find clean and corrupt pairs.

Previous works have shown vision transformers can correctly identify objects in various contexts regardless of the foreground or background (Vilas et al., 2024). Since image encoders in multimodal models, such as BLIP and LLaVA, are robust to minor variations to the context of an object, only the semantic contents of the image need to be controlled to obtain reliable patching results (Vilas et al., 2024). In contrast, masking pixels, adding Gaussian noise to image patches, or replacing pixels would create unnatural images, leading to unreliable results. In Section 4.1, we demonstrate the effectiveness of SIP corruption over Gaussian-noise image corruption and experiment with using generative models to create SIP.

NOTICE VQA Task Design. We adapt the SVO-Probes, MIT States, and Facial Expressions image-classification datasets into VQA tasks. We prompt models to choose between a correct answer that matches the image and an incorrect, semantically different alternative, as seen in Figure 1. For example, using the image-class labels “happy” and “angry,” we create a VQA prompt: “Is this person feeling happy or angry?”. For LLaVA, we adapt the standard prompting format to ensure the model selects one of the multiple choice options: “USER: <image> [Prompt]. Answer with one word. Assistant:”. While in SVO-Probes, each subject, verb, or object comes with its own semantically different counterpart, we constructed such pairs for MIT-States and Facial Expression. In MIT-States, we separated states into color, shape, and texture and selected counterparts from the respective categories for each question. We grouped similar states (e.g., “huge”, “large”) together and ensured that the

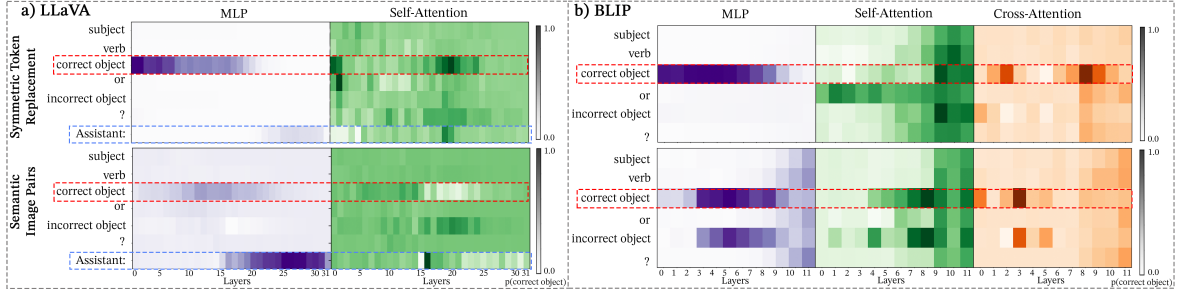


Figure 3: Module-wise activation patching results for BLIP and LLaVA on “objects” from SVO-Probes. We visualize the restoration probability after patching for MLP, self-attention, and cross-attention layers in the image-grounded text-encoder. The y-axis denotes which token we patch, and the x-axis denotes which layer we patch.

selected counterpart would not be similar in meaning to the correct answer (using cosine similarity and word2vec word embeddings (Mikolov et al., 2013)). Similarly, in the Facial Expressions dataset, we grouped the seven emotions into positive, negative, and neutral and selected counterparts from opposing emotions. We include more details on the VQA design and the datasets in Appendix A and Appendix B, respectively.

Activation Patching. Let M be a vision-language model and M_l its hidden states at layer $l \in \{0, 1, \dots, n\}$. The forward pass on an image (I) and text (T) pair at layer l is denoted as $M_l(I, T)$. We define I^* as a corrupted image and T^* as a corrupted text input. Note that we only corrupt one modality at a time. In a forward pass with I^* and T , activation patching replaces the hidden states of the corrupted run $M_l(I^*, T)$ with the clean run $M_l(I, T)$ to determine which components have the most significant indirect effect on the output. The hidden states at all subsequent layers, $M'_j(I^*, T)$ for $j \in \{l + 1, \dots, n\}$, are called *patched states*. In Figure 2, we provide an example illustrating patching using SIP image corruption. The image of the puppy is the clean image, I , and the goat is I^* . Patching the correct answer token “puppy” at $M_l(I, T)$, from the clean image to the “puppy” token at $M(I^*, T)$ creates the patched states $M'(I^*, T)$ shown as orange diamonds in Figure 2.

Measuring Patching Effects. The most common ways of measuring patching effect are *restoration probability* and *logit difference*. Let L^* and L' denote the logits from the corrupted and patched runs, and let τ, τ^{inc} denote the correct/incorrect multiple-choice answer. Logit difference measures the indirect impact patching has on the model’s logits. Let $L(\tau, \tau^{\text{inc}}) = \text{Logit}(\tau) - \text{Logit}(\tau^{\text{inc}})$. The logit difference is defined as $L'(\tau, \tau^{\text{inc}}) - L^*(\tau, \tau^{\text{inc}})$.

Restoration probability measures the impact of patching clean states into a corrupt run on predicting the correct answer token: $P'(\tau) - P^*(\tau)$.

4 Results

We begin our investigation by conducting module-wise activation patching to understand the global patterns of self-attention, cross-attention, and MLP modules under both image and text corruption. We demonstrate that SIP image corruption produces results closely aligned with STR text corruption and effectively highlights the role of middle-layer MLPs, which have repeatedly proven crucial in a wide range of tasks (Meng et al., 2023). In contrast, Gaussian Noise image corruption only identifies the final MLP layers as producing a significant patching effect. Building on this, we use SIP to perform a fine-grained analysis of cross-attention heads in BLIP and self-attention heads in LLaVA, the critical point of multimodal integration for each model. As we identify the most significant cross-attention heads in BLIP and most significant self-attention heads in LLaVA, we categorize them into three distinct functional roles, offering deeper insights into their contributions and behaviors.

4.1 Module-wise Activation Patching

Module-wise activation patching reveals how different model components (i.e., attention blocks and MLPs) of VLMs respond to image and text corruption, highlighting which layers and modules (MLPs, self-attention, cross-attention) are most influential for successfully completing VQA tasks. First, we compare the effects of STR and SIP on BLIP and LLaVA, as shown in Figure 3 and Appendix F.2. Across all tasks, we observe: (1) MLP patching reveals that image corruption emphasizes the importance of middle layers, while text corruption primarily affects earlier to middle layers;

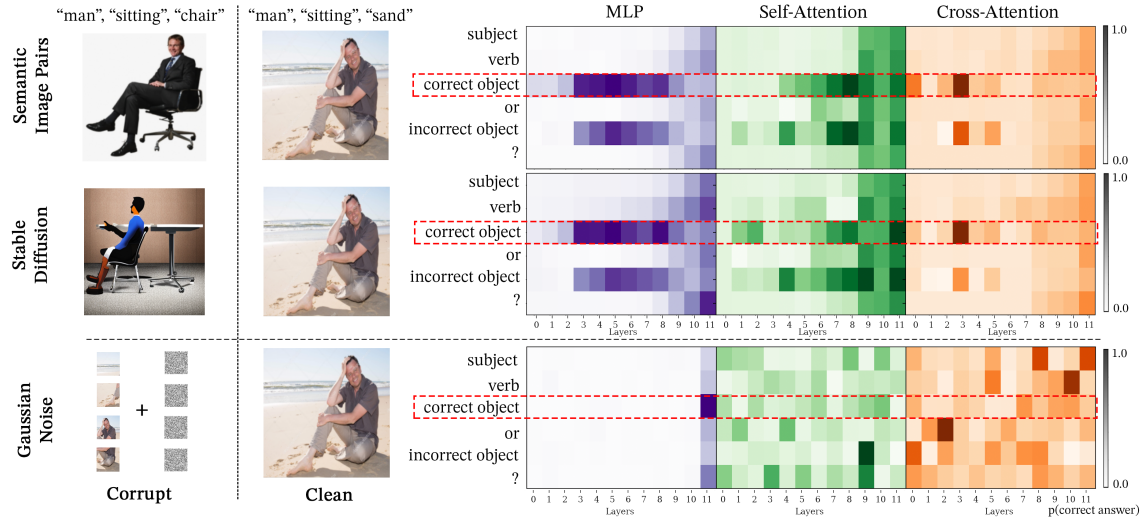


Figure 4: Module-wise activation patching results for SIP and Gaussian-Noice corruption on SVO-Probes on BLIP. SIP corruption produces activation patterns that align with Stable Diffusion results and highlight the importance of middle layers, while Gaussian noise fails to reveal meaningful attention layers, emphasizing the effectiveness of SIP for probing vision-language models.

(2) Under text corruption, LLaVA’s self-attention demonstrates strong localization, on the correct object token in layer 0 and layer 21, similar to the cross-attention localization on the correct answer seen in BLIP at layer 3; (3) Under image corruption, LLaVA’s self-attention attends away from the correct answer token in layers 16-25; and (4) LLaVA’s earlier layers identify the correct answer token while in the later layers the emphasis shift to the instruction token “Assistant:”.

We evaluate the effectiveness of SIP corruption by comparing it with Gaussian-noise, the image corruption technique used in previous work (Palit et al., 2023). In addition, we explore semantic image corruption using generative models, which shows the NOTICE framework is generalizable beyond curated datasets like SVO-Probes. Specifically, we use in-painting Stable Diffusion (Rombach et al., 2022) to perform semantic image corruption by generating “corrupted” versions of clean images that only vary on one key semantic property. For instance, generating an image of a man sitting on a chair from an image of a man sitting on sand. Figure 4 demonstrates the promise of using generative models to implement SIP corruption since stable-diffusion module-wise patching results are highly similar to SVO-Probes. Note that generative models may occasionally introduce undesirable image artifacts (e.g., adding extra fingers to a hand) that we cannot control. We further analyze this in Appendix D. While using Stable Diffusion to create SIP shows promise for future

CMA applications, Gaussian noise proves unreliable. In both LLaVA and BLIP, MLP modules highlight the importance of early and middle layers, while Gaussian noise only highlights the last layer (see Figure 4 and Appendix F.1). Several studies highlight the importance of middle layers in transformers (Durrani et al., 2020; Meng et al., 2023), and show that MLPs store syntactic information and task-specific representations often peak in middle layer MLPs (Durrani et al., 2020; Hewitt and Manning, 2019; Goldberg, 2019; Jawahar et al., 2019). Additionally, self-attention patching in SIP assigns the highest probability to the correct token, unlike Gaussian noise. SIP and STR also display consistent activation patterns—long horizontal regions in MLPs (aligning with (Meng et al., 2023)) and vertical regions in self-attention—whereas Gaussian Noise shows clear differences. These findings suggest that SIP is a more effective tool for probing VLMs and identifying key components and interactions.

4.2 Universal Attention Heads Exist Across Corruption Schemes and Tasks

Following the module-wise analysis, we focus on the attention heads, investigating their role in facilitating vision-language integration. We calculate the logit difference for each cross-attention head in BLIP and self-attention head in LLaVA by analyzing the impact of patching the correct answer across the SVO-Probes, MIT States, and Facial Expressions datasets in BLIP and the “As-

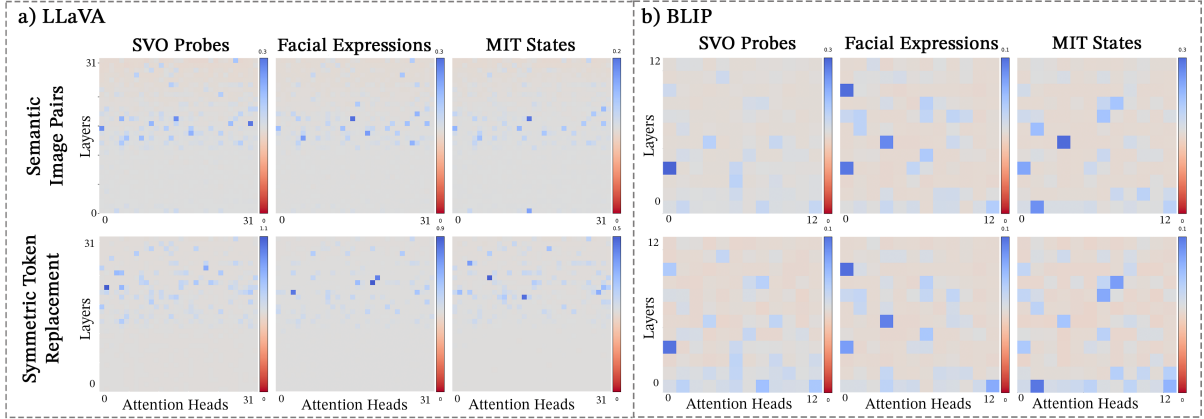


Figure 5: Logit difference demonstrating the impact of patching the correct answer for each LLaVA self-attention head and BLIP cross-attention head on the SVO-Probes, MIT States, and the Facial Expressions datasets. Many key attention heads overlap in importance across both modalities.

Model	Vision Heads	Multimodal Heads	Text Heads
LLaVA	L15.H5, L16.H18 L14.H4, L15.H8 L18.H26, L17.H0 L17.H22, L14.H28 L19.H4, L15.H14 L17.H13, L20.H28	L28.H7, L31.H27 L16.H24, L16.H24 L19.H15, L18.H30 L18.10, L21.H30	L22.H16, L29.H19
BLIP	L5.H3	L3.H0	L0.H11

Table 1: Universal attention heads in LLaVA and BLIP, identified as having consistently high logit differences across tasks and modality corruptions. In LLaVA, most critical self-attention heads serve multimodal or vision-specific roles. In contrast, BLIP’s dual-stream design results in distinct vision, text, and multimodal heads, supporting more modular cross-modal integration.

stantant:” token in LLaVA . We perform activation patching on the “Assistant:” token since Figure 3 demonstrates that patching the “Assistant:” creates the largest module-wise patching effect. See Appendix G for more details, as well as for the patching of the correct answer token in each self-attention head in LLaVA. Figure 5 demonstrates that, in BLIP, image corruption has a greater impact on logit difference compared to text corruption in SVO-Probes and MIT-States. Specifically, image corruption results in up to 20% logit difference, while text corruption causes a difference of 6%, suggesting BLIP relies more on visual input than text for predicting the correct answer. In LLaVA, Figure 5 shows the logit difference is consistent for both modalities. These results contrast with previous findings that VLMs are more influenced by textual priors (Salin et al., 2022). Despite the difference in scale in BLIP, many critical attention heads identified under text corruption are also crucial when images are corrupted, with this overlap

persisting across tasks. To quantify the overlap between the two modalities’ corruptions and across tasks, we identify three universal cross-attention heads in BLIP (L0.H11, L3.H0, and L5.H3) and seven universal multimodal self-attention heads in LLaVA (L28.H2, L31.H27, L18.H10, L16.H24, L21.H30, L19.H15, and L18.H30) where $Lx.Hy$ denotes layer x , head y .

Table 1 shows “universal attention heads”, which we defined as consistently having a logit difference two standard deviations above the average regardless of the task or the modality corruption. In LLaVA, we find only find two “text only” heads; a majority of important self-attention heads either overlap with multimodal heads that are important for both modalities or emerge under image corruption (“vision only” heads). This suggests visual inputs still influence the prompt, even under text corruption. We hypothesize this is due to LLaVA’s early fusion, where image tokens are embedded into the textual prompt. In contrast, BLIP has exactly one vision, one text, and one multimodal head, likely because its dual-stream cross-attention design allows the model to fuse visual and textual information more effectively without requiring multiple specialized heads for different modalities. These findings highlight key differences in cross-modal integration between the two models.

The presence of overlapping attention heads when patching different modalities and tasks is a novel discovery. Since cross-attention heads in BLIP and self-attention heads in LLaVA integrate vision and language, overlapping heads suggest that vision and language convey similar information to the VLM despite being distinct feature

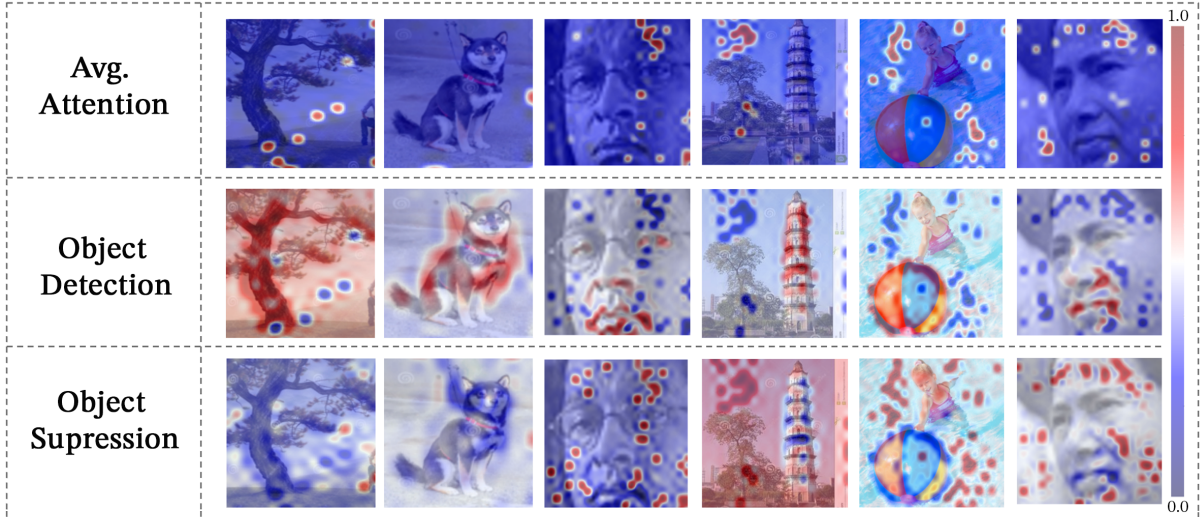


Figure 6: Universal cross-attention heads in BLIP implement object detection and object suppression. We visualize the average of all attention heads compared to the universal heads performing each function.

spaces. This challenges the longstanding notion that text is the dominant modality for VLMs. Additionally, the reuse of these components across tasks offers insights into model behavior and model editing, enabling targeted improvements that enhance performance, address bias, and better adapt the model to new tasks.

4.3 Universal Attention Heads Perform Distinct Functions

To better understand the role of universal attention heads in multimodal integration, we visualize the cross-attention patterns between the “correct answer” text token and image patches in BLIP. For LLaVA, we visualize self-attention patterns of universal multimodal self-attention heads from the “Assistant:” token to image patches. We find that universal cross-attention heads in BLIP fall into three function classes: *object detection*, *object suppression*, and *outlier suppression*, seen in Figure 6 whereas universal multimodal self-attention heads in LLaVA primarily implement outlier suppression seen in Figure 7.

Outliers in Attention Patterns. Similar to Darcet et al. (2024), we find that cross-attention patterns in BLIP and text-to-image self-attention patterns in LLaVA are dominated by “outlier features” with extremely high average attention scores. These outliers appear in Figure 6 as non-informative image patches with dark red scores. Unlike in ViTs, where outliers are thought to obscure interpretable features in vision attention, we

find attention functions reduce attention to these outliers. In BLIP, this enables object detection by linking text tokens to the relevant image patches. Our finding that outliers play an important role in VLM representations is similar to recent works showing outliers in LLMs store task-specific information (Rudman et al., 2023).

Object Detection. In BLIP, cross-attention layer 3, head 0 (L3.H0) implements object detection by linking a text token with relevant image patches. These heads associate physical objects in images with the nouns in SVO-Probes and MIT States. Notably, L3.H0 can segment images even when the word does not explicitly appear in the image. For example, in the Facial Expressions dataset, L3.H0 implicitly associates the correct emotion (e.g., “anger”) with the mouth region. L3.H0 is the only universal cross-attention head identified as “multimodal,” consistently demonstrating its importance across all tasks, regardless of modality corruption. Its ability to map text tokens to image patches highlights its critical role in fusing vision-language information and demonstrates its true multimodal capabilities.

Object Suppression. Cross-attention head L5.H3 in BLIP explicitly attends *away* from the image patches associated with an input token and attends strongly to the outlier cross-attention patches. Object suppression cross-attention heads perform the inverse of object detectors, as they explicitly reduce attention to specific objects in an image. In SVO-Probes, this amounts to attending away

from subjects or objects, and in MIT-States, L5.H3 attends away from relevant portions of the image that form the noun that is changing states. However, L5.H3 does not perform object suppression in Facial Expressions. Instead, L5.H3 performs the closely related function of outlier suppression. L5.H3 in BLIP provides a rare example of an *interpretable polysemantic* attention head that implements different interpretable functions depending on the input (Gould et al., 2023).

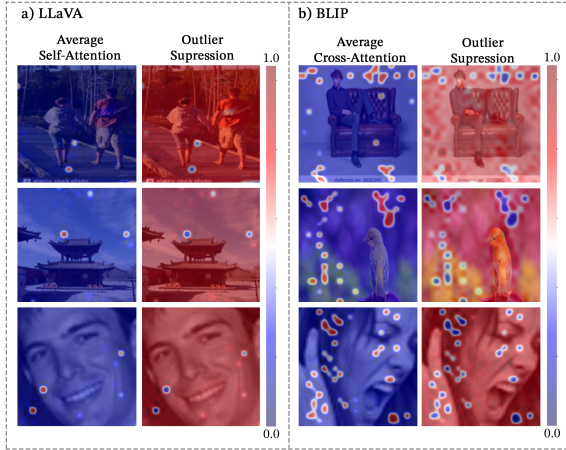


Figure 7: Universal heads in BLIP and LLaVA perform outlier suppression. We visualize the average of all attention heads compared to the universal heads performing outlier suppression.

Outlier Suppression. Outlier suppression heads reduce attention to high-probability image patches and distribute attention uniformly across remaining patches. All universal multimodal attention heads in LLaVA implement outlier suppression. In BLIP, L0.H11 implements outlier suppression across all tasks in this paper. For BLIP, L0.H11 is a text-only universal cross-attention head, meaning it does not produce a significant patching effect when images are corrupted. Due to its specialization, L0.H11 does not contribute to the semantic association between text tokens and image patches. Instead, it plays a crucial role in filtering out irrelevant visual information, enhancing the model’s robustness.

5 Conclusion

In this work, we study the roles of cross-attention and self-attention in multimodal integration. To accomplish this, we introduce NOTICE, a Gaussian-Noise-free Text-Image Corruption and Evaluation pipeline for mechanistic interpretability in VLMs. Using our SIP image corruption framework and STR text corruption, NOTICE enables causal me-

diation analysis across both modalities. Experiments across three tasks reveal “universal attention heads” with consistent patching effects across BLIP and LLaVA. We show that cross-attention supports three functions: object detection, object suppression, and outlier suppression. However, self-attention primarily handles outlier suppression, indicating cross-attention’s unique role in image grounding. While attention presents a plausible, but not necessarily faithful rationale for model prediction (Wiegrefe and Pinter, 2019), our findings show patterns of attention heads that consistently, and causally, influence model logits. With the support NOTICE, applications using VLMs can benefit from more transparent and adaptable models, facilitating deeper insights and enhanced performance.

6 Limitations

The NOTICE framework, combined with SIP corruption, offers valuable insights into the interpretability of vision-language models. In this paper, we applied our dual corruption scheme to three diverse datasets and experimented with generative models for image corruption. However, many more datasets remain unexplored. For instance, shape-based datasets like ShapeWorld (Kuhle and Copestake, 2017) and CLEVR (Johnson et al., 2017) allow variations in color and shape for corruption. Similarly, icon datasets like IconQA (Lu et al., 2021) and Emojis (Kralj Novak et al., 2015), as well as specialized datasets such as DeepFashion (Liu et al., 2016), Caltech-UCSD Birds-200-2011 (Wah et al., 2011), and the Stanford Cars dataset (Kramberger and Potočník, 2020), can be turned into “clean” and “corrupt” pairs. In future work, we will expand our experiments to include a wider range of datasets and models, further enhancing the robustness and generalizability of our findings.

References

- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.
- Yonatan Belinkov. 2022. [Probing classifiers: Promises, shortcomings, and advances](#). *Computational Linguistics*, 48(1):207–219.
- Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. 2024. [Impossibility theorems for feature attribution](#). *Proceedings of the National Academy of Sciences*, 121(2).

- 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](#). In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI*, page 565–580, Berlin, Heidelberg. Springer-Verlag.
- Bilal Chughtai, Lawrence Chan, and Neel Nanda. 2023. [A toy model of universality: Reverse engineering how networks learn group operations](#). *Preprint*, arXiv:2302.03025.
- Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. [What you can cram into a single \$\\$&!#*\$ vector: Probing sentence embeddings for linguistic properties](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.
- Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. 2024. [Vision transformers need registers](#). *Preprint*, arXiv:2309.16588.
- Beddiar Djamil-Romaissa, Oussalah Mourad, and Sepänen Tapio. 2022. Automatic captioning for medical imaging (mic): a rapid review of literature.
- Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. 2020. Analyzing individual neurons in pre-trained language models. *arXiv preprint arXiv:2010.02695*.
- Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. *arXiv preprint arXiv:2109.04448*.
- Mario Giulianelli, Jacqueline Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2021. [Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information](#). *Preprint*, arXiv:1808.08079.
- Yoav Goldberg. 2019. Assessing bert’s syntactic abilities. *arXiv preprint arXiv:1901.05287*.
- Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. 2013. Challenges in representation learning: A report on three machine learning contests. In *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part III* 20, pages 117–124. Springer.
- Rhys Gould, Euan Ong, George Ogden, and Arthur Conmy. 2023. [Successor heads: Recurring, interpretable attention heads in the wild](#). *Preprint*, arXiv:2312.09230.
- Songhao Han, Le Zhuo, Yue Liao, and Si Liu. 2023. Llms as visual explainers: Advancing image classification with evolving visual descriptions. *arXiv preprint arXiv:2311.11904*.
- Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. [How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model](#). *Preprint*, arXiv:2305.00586.
- Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*.
- Jack Hessel and Lillian Lee. 2020. [Does my multimodal model learn cross-modal interactions? it’s harder to tell than you might think!](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 861–877, Online. Association for Computational Linguistics.
- John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.
- Nikolai Ilinykh and Simon Dobnik. 2022a. [Attention as grounding: Exploring textual and cross-modal attention on entities and relations in language-and-vision transformer](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4062–4073, Dublin, Ireland. Association for Computational Linguistics.
- Nikolai Ilinykh and Simon Dobnik. 2022b. Attention as grounding: Exploring textual and cross-modal attention on entities and relations in language-and-vision transformer. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4062–4073.
- Phillip Isola, Joseph J Lim, and Edward H Adelson. 2015. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1383–1391.
- Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.
- Nick Jiang, Anish Kachinthaya, Suzie Petryk, and Yossi Gandelsman. 2024. Interpreting and editing vision-language representations to mitigate hallucinations. *arXiv preprint arXiv:2410.02762*.
- Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

- Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of emojis. *PloS one*, 10(12):e0144296.
- Tin Kramberger and Božidar Potočnik. 2020. Lsun-stanford car dataset: enhancing large-scale car image datasets using deep learning for usage in gan training. *Applied Sciences*, 10(14):4913.
- Alexander Kuhnle and Ann Copestake. 2017. Shapeworld-a new test methodology for multimodal language understanding. *arXiv preprint arXiv:1704.04517*.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. [Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation](#). *Preprint*, arXiv:2201.12086.
- Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.
- Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer.
- Paul Pu Liang, Yiwei Lyu, Gunjan Chhablani, Nihal Jain, Zihao Deng, Xingbo Wang, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2023. [Multiviz: Towards visualizing and understanding multimodal models](#). *Preprint*, arXiv:2207.00056.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. Large-scale fashion (deepfashion) database. *The Chinese University of Hong Kong, Category and Attribute Prediction Benchmark, Xiaoou TangMultimedia Laboratory*.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.
- Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*.
- Yujie Lu, Xiujun Li, William Yang Wang, and Yejin Choi. 2023. [Vim: Probing multimodal large language models for visual embedded instruction following](#). *Preprint*, arXiv:2311.17647.
- Grace Luo, Trevor Darrell, and Amir Bar. 2024. Task vectors are cross-modal. *arXiv preprint arXiv:2410.22330*.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. [Locating and editing factual associations in gpt](#). *Preprint*, arXiv:2202.05262.
- Sachit Menon and Carl Vondrick. 2022. Visual classification via description from large language models. *arXiv preprint arXiv:2210.07183*.
- Luke Merrick and Ankur Taly. 2020. The explanation game: Explaining machine learning models using shapley values. In *Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4*, pages 17–38. Springer.
- Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2024. [Circuit component reuse across tasks in transformer language models](#). *Preprint*, arXiv:2310.08744.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. [Efficient estimation of word representations in vector space](#). *Preprint*, arXiv:1301.3781.
- Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. 2024. Towards interpreting visual information processing in vision-language models. *arXiv preprint arXiv:2410.07149*.
- Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. 2023. Towards vision-language mechanistic interpretability: A causal tracing tool for blip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2856–2861.
- Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. 2020. Seeing past words: Testing the cross-modal capabilities of pretrained v&l models on counting tasks. *arXiv preprint arXiv:2012.12352*.
- 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.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.
- William Rudman, Catherine Chen, and Carsten Eickhoff. 2023. [Outlier dimensions encode task specific knowledge](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14596–14605, Singapore. Association for Computational Linguistics.
- Emmanuelle Salin, Badreddine Farah, Stéphane Ayache, and Benoit Favre. 2022. [Are vision-language transformers learning multimodal representations? a probing perspective](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:11248–11257.

- Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017a. [FOIL it! find one mismatch between image and language caption](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, Vancouver, Canada. Association for Computational Linguistics.
- Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017b. Foil it! find one mismatch between image and language caption. *arXiv preprint arXiv:1705.01359*.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. 2022. [Denoising diffusion implicit models](#). *Preprint*, arXiv:2010.02502.
- Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.
- Martina G Vilas, Timothy Schaumlöffel, and Gemma Roig. 2024. Analyzing vision transformers for image classification in class embedding space. *Advances in neural information processing systems*, 36.
- Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset.
- Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. [Interpretability in the wild: a circuit for indirect object identification in gpt-2 small](#). *Preprint*, arXiv:2211.00593.
- Sarah Wiegrefe and Yuval Pinter. 2019. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*.
- Peng Xu, Xiatian Zhu, and David A. Clifton. 2023. [Multimodal learning with transformers: A survey](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132.
- Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. 2023. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10714–10726.
- Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*.
- Fred Zhang and Neel Nanda. 2024. [Towards best practices of activation patching in language models: Metrics and methods](#). *Preprint*, arXiv:2309.16042.
- Pengchuan Zhang, Xiujuan Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588.

A VQA-Prompts

For each task, we create a VQA prompt that asks the model to choose between a correct answer, which corresponds to the image, and an incorrect answer that corresponds to a semantically minimal pair. While in SVO-Probes, each subject, verb, or object, came with its own semantically different counterpart, in MIT-States dataset and Facial Expression dataset, we constructed such counterparts manually. In MIT-States, we separated states into color, shape, and texture, and selected counterparts for each question from the respective categories. We grouped similar states (e.g., “huge”, “large”) together, and ensure that the selected counterpart would not be similar in meaning to the correct answer. Similarly, in the Facial Expressions dataset, we grouped the seven emotions into positive, negative, and neutral, and selected counterparts from opposing emotions, to ensure the model can correctly chose one answer. Recall that the SVO-Probes dataset consists of subject, verb, and object triplets. Each triplet is presented in singular form and non-conjugated. We use GPT-3.5 Turbo to perform grammatical error correction on all input prompts in SVO-Probes while minimally perturbing the template structure. We evaluate BLIP using exact match accuracy, which requires the generated answer to match the correct answer exactly, including capitalization and conjugation. BLIP achieves 66.1% accuracy on the facial expressions dataset, 44.6% on MIT States, and 16.1% on SVO-Probes (see Table 2 for category-wise accuracy). In our experiments, we use samples that BLIP accurately predicted. For each task, we select 500 samples across all categories: subjects, verbs, and objects for SVO-probes; colors, shapes, and textures for MIT-States; and one of the five emotions BLIP correctly identified for the facial expression identification task. To mitigate model bias toward the position of the correct answer in multiple-choice prompts, we equally distribute 250 samples where the correct option precedes and 250 where it follows the “or” token.

B Datasets

SVO-Probes. The Subject-Verb-Object (SVO) Probes dataset (Hendricks and Nematzadeh, 2021) consists of over 48,000 image-language pairs that vary on exactly one of the subject, verb, and object tuples that describe an image. For example, Figure 1 depicts image-text pairs with the captions “A

child crossing the street” (child, crossing, street) and “A lady crossing the street” (lady, crossing, street). In this instance, the two image pairs only vary on the subject of the image, but the verb (crossing) and the object (street) are identical. SVO-Probes tests if multimodal models can capture fine-grained linguistic information about the content of an image.

MIT-States. The MIT States dataset (Isola et al., 2015) is a collection of 245 object classes, 115 attribute classes, and over 53,000 images. The dataset is used to study how models can generalize state transformations across different object classes. For instance, learning what “melted” looks like on butter can help recognize “melted” in chocolate.

Facial Expressions. The Facial Expression Recognition 2013 (FER-2013) (Goodfellow et al., 2013) dataset includes 35,887 grayscale images of faces resized to 48x48 pixels. The images are categorized into seven emotion classes: anger, disgust, fear, happiness, sadness, surprise, and neutral. The FER-2013 dataset is valuable for comparing feature learning methods with hand-engineered features in emotion recognition tasks.

C BLIP and LLaVA Performance

Dataset	BLIP Accuracy (%)	LLaVA Accuracy (%)
SVO Probes	16.10	40.89
MIT States	44.60	67.81
Facial Expressions	66.10	79.31

Table 2: Comparison of BLIP and LLaVA overall accuracy by dataset, expressed as percentages.

In evaluating the BLIP and LLaVA, we focused on exact match accuracy, which necessitates that the model’s generated answer must exactly match the correct answer in terms of both content and form, including specifics like capitalization and conjugation. This stringent criterion ensures that the model’s understanding and response generation align closely with the intended semantics of the test prompts. Table 2 shows the performance of BLIP and LLaVA (zero-shot) on each task, and each category within the task.

D SIP with Stable Diffusion

With the primary goal of establishing a more controlled environment conducive to the generation of semantically similar corrupt (negative) images

from clean(positive) images, we also explore the use of stable diffusion to produce images that are similar to clean images from the SVO-Probes dataset, with the sole variation being in only one among the subject, verb, or object of the image.

Our experiment design utilizes the *stable-diffusion-2-inpainting* (Rombach et al., 2022) to generate a corrupt image using the clean image as a base template. In this context, we define a prompt structure for the model as “a picture of $\text{sub}_{neg} + \text{verb}_{neg} + \text{obj}_{neg}$ ” where $\langle \text{sub}_{neg}, \text{verb}_{neg}, \text{obj}_{neg} \rangle$ denote the svo-triplet for the negative image.

For the parametric constraints on the diffusion model pipeline, we utilize a DDIM Scheduler (Song et al., 2022), setting the DDIM value = 120 and the prompt guidance scale value to 8 for all the image generations. For generating images on the template of the base positive image, we use a contour mask, which allows for inpainting of the negative subject, verb, or object into the positive image. In such cases, a variation in quality is observable in the images introduced due to the probabilistic nature of the model.

Holistically, the generated images from *stable-diffusion-2-inpainting* are largely closer to the original corrupt images from SVO-Probes dataset where the image annotations utilize simpler terminologies to depict any one of the three $\text{sub}_{neg}, \text{verb}_{neg}, \text{obj}_{neg}$ such as “grass” instead of “meadow” (refer to Fig. 8 and 9).

E Compute Resources

For each experiment, we use one NVIDIA GeForce RTX 3090 GPU. Running BLIP for each activation patching experiment ranged from 12 hours for SVO probes to 10 hours in MIT states and Facial Expressions. Running LLaVA for each activation patching experiment ranged from 96 hours for SVO probes to 90 hours in MIT states and Facial Expressions.

F Results

F.1 SIP vs. Gaussian Noise Image Corruption.

Figure 10 illustrates the comparative effects of Semantic Image Pairs (SIP) and Gaussian Noise (GN) image corruption across all tasks, highlighting SIP’s superior ability to emphasize the significance of middle layers in the MLP, in contrast to GN’s focus on later layers. These observations affirm SIP’s enhanced reliability and effectiveness

in revealing the nuanced functioning of vision-language models, supporting its use for in-depth interpretability studies.

F.2 Module-wise Activation Patching

Module-wise activation patching offers a detailed examination of how different components of the BLIP model respond to image and text corruption, using Semantic Image Pairs (SIP) and Gaussian noise for images and Symmetric Token Replacement (STR) for texts. This approach is crucial for uncovering the roles of self-attention, cross-attention, and MLP modules within the architecture, particularly focusing on their reaction to varied corruption methods. Through these experiments, we aim to pinpoint the layers most critical for the model’s decision-making processes and explore how different corruption strategies affect the transparency and reliability of the model’s internal workings. This analysis not only tests the robustness of the SIP technique against the traditional Gaussian noise approach but also highlights the strategic layers that are pivotal in handling multi-modal data integration.

First, Figure 11 and Figure 12 show module-wise activation patching for SVO Probes, broken down into subjects, verbs, and objects, VQA prompts.

Next, we examine module-wise activation patching for the Facial Expressions dataset in Figure 13 for BLIP and Figure 14. We note that image and text corruption results in similar patterns. This provides further support to SIP corruption aligning with STR corruption.

Lastly, Figure 15 and Figure 16 exhibit outcomes similar to the other tasks but with a notable distinction: there’s an increased likelihood of accurately predicting the correct token within the cross-attention heatmaps during text corruption. This suggests that the multiple-choice VQA prompts used in the MIT States dataset may be more definitive compared to the other datasets. While distinguishing between emotions such as “angry” and “sad,” or identifying subtle differences like “woman” versus “girl” can be challenging in the Facial Expressions and SVO Probes datasets, respectively, the task of differentiating between clearly contrasting attributes like “modern” versus “ancient” buildings in MIT States appears to be more straightforward.

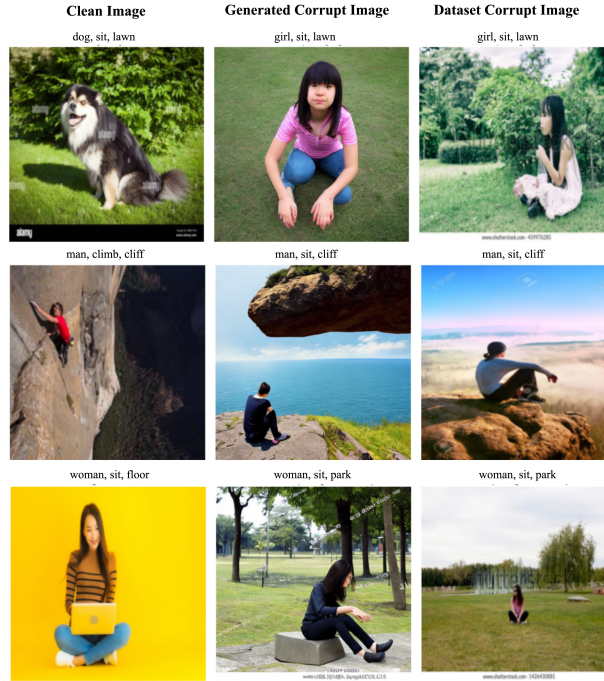


Figure 8: Successful examples of images generating using *stable-diffusion-2-inpainting* that showcase corrupt images close enough to the clean image yet different in (1) Subject (2) Verb (3) Object

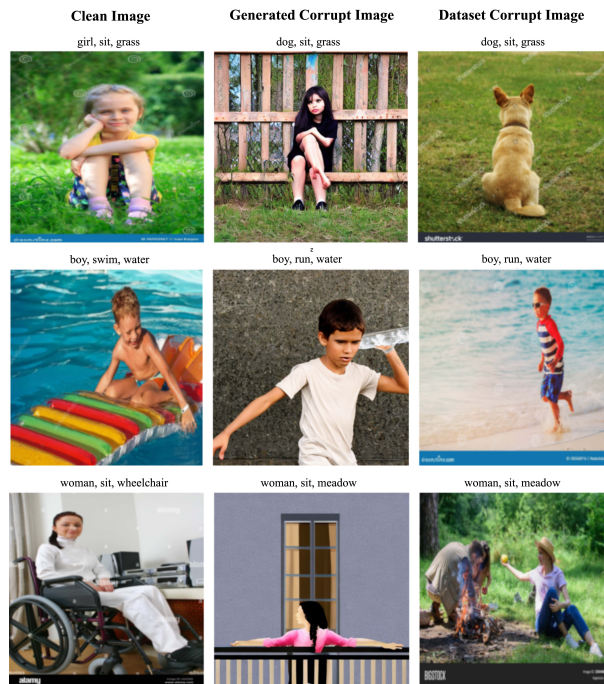


Figure 9: Poor examples of images generating using *stable-diffusion-2-inpainting* that showcase corrupt images which deviate from the clean image in (1) Subject (2) Verb (3) Object

G LLaVA Self-Attention Patching Details

We consider two methods for performing activation on LLaVA’s self-attention heads. First, we patch only on the correct answer token. Although Figure 5 demonstrates a strong patching effect in BLIP, patching the correct answer token in LLaVA

results in an extremely small logit difference, with the largest average logit difference being 0.016 as shown in Figure G, regardless of modality corruption or dataset. Patching the correct answer token does not result in any “universal attention heads” indicating that self-attention does not play a pivotal

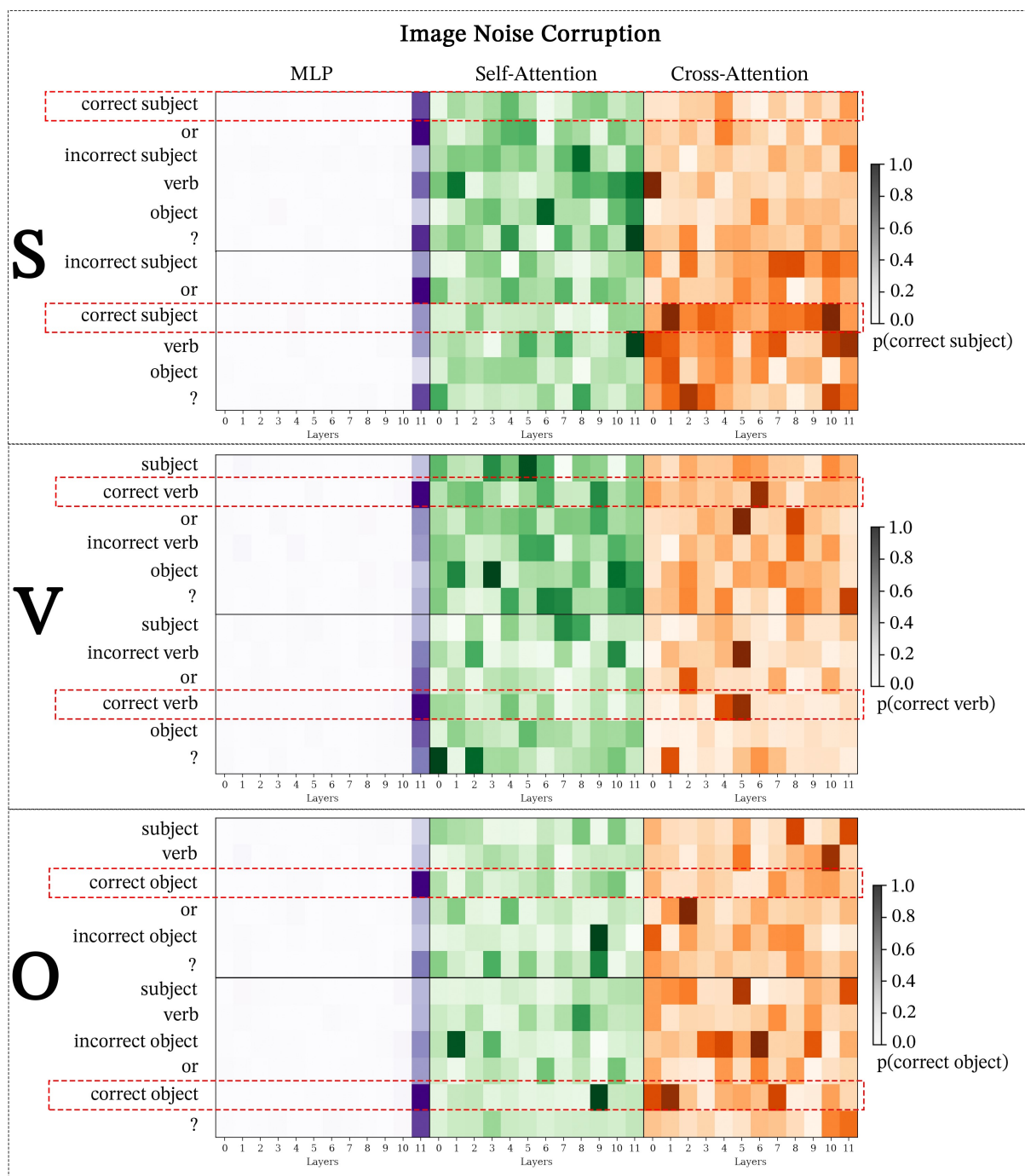


Figure 10: Comprehensive view of Gaussian noise effects across all tasks in the SVO-Probes dataset, providing a contrast to SIP effects and highlighting the differential impact of noise on model accuracy and layer activation.

Cross-Attention Functions	SVO	MIT	Facial Expressions
object detection	L3.H0	L3.H0	L3.H0
object suppression	L5.H3	L5.H3	N/A
outlier suppression	L0.H11	L0.H11	L0.H11, L5.H3

Table 3: Universal heads identified through cross-attention head patching, alongside their functions.

role in associating the relevant images patches with the correct answer token in LLaVA. Module-wise

patching using image corruption for both MLPs and Self-Attention blocks shows that patching the 'Assistant:' token consistently produces the strongest patching effect for LLaVA. Accordingly, we experiment with patching the 'Assistant:' token for LLaVA's self-attention heads. Patching the "Assistant:" token creates a far stronger logit difference with some heads producing an average logit difference as high as 0.4. Patching 'Assistant:' produces

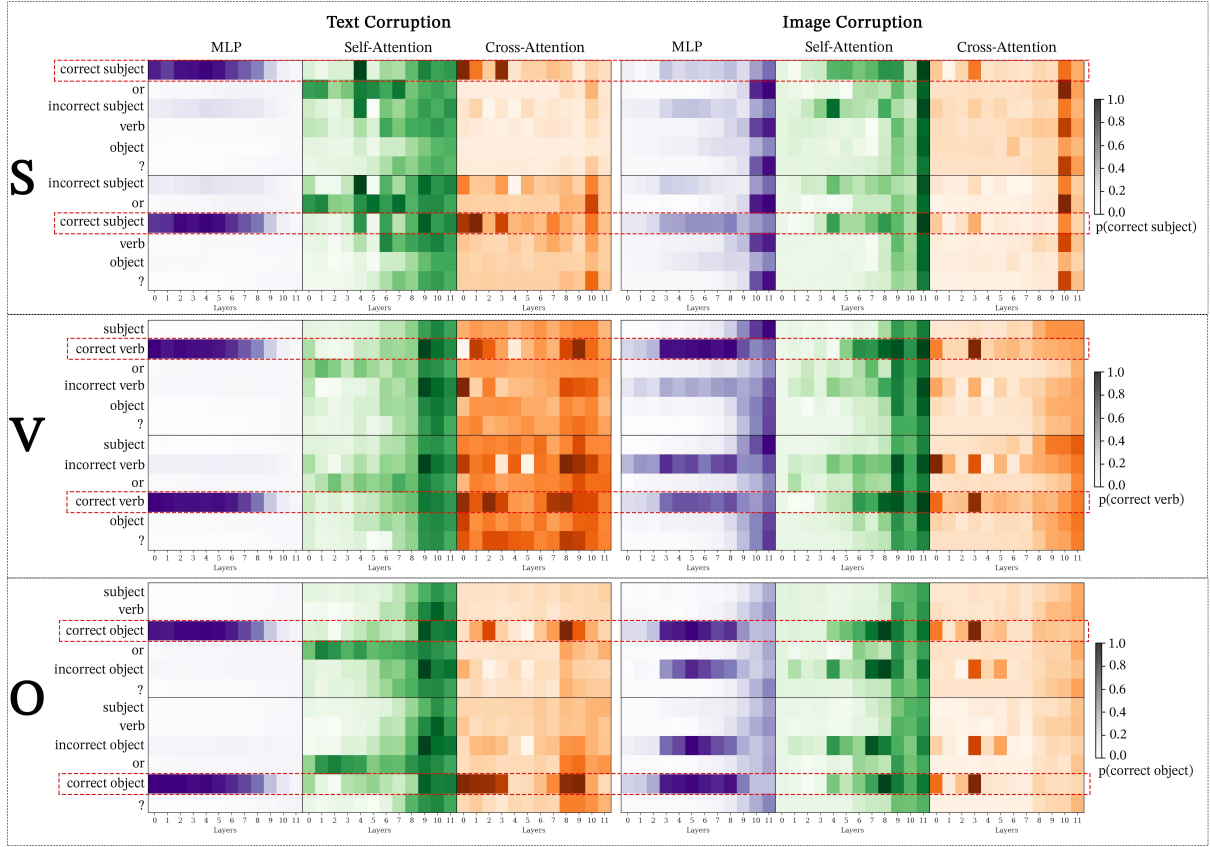


Figure 11: Module-wise activation patching results on BLIP for the SVO-Probes dataset.

consistent results with 6 universal self-attention heads compared to 0 when patching the correct answer token. Both the consistency and strength of patching the 'Assistant:' token highlight the importance of the 'Assistant:' token in storing information relevant to the image. Future studies performing activation patching on instruction-tuned VLMs that utilize causal self-attention to fuse modalities should focus on the information in the specific instruction tokens.

G.1 Comparing Cross-Attention and Self-Attention Head Functions

Our result that the cross-attention mechanism VLMs consistently associates the nouns with their corresponding image patches and that causal self-attention does not associate nouns to image regions is supported by previous works investigating cross-attention and self-attention patterns in toy VLMs (Ilinykh and Dobnik, 2022a). Although self-attention heads in LLaVA associate neither the correct answer token nor the 'Assistant:' token with the relevant image patches, universal self-attention heads in LLaVA have the primary function of mitigating outlier attention patterns in images.

Important cross-attention heads in BLIP are scattered throughout the network ranging from layer 0 to layer 9. In contrast, import self-attention heads in LLaVA are concentrated in the middle layers of the network with x / y self-attention heads between 14-18.

G.2 Cross-Attention Head Function Classes

Understanding the specific functions of cross-attention heads within vision-language models (VLMs) is crucial for mechanistic interpretability. In this section, we delve into the distinct roles played by universal cross-attention heads identified by our NOTICE scheme. These heads can be categorized into three primary function classes: *implicit object detection*, *object suppression*, and *outlier suppression*. Below, we provide detailed examples of these functions along with corresponding visualizations from the SVO-Probes, MIT-States, and Facial Expressions datasets.

L0.H11 - outlier suppression. Universal cross-attention head L0.H11 implements outlier suppression in all tasks. It actively attends away from high-probability image patches and uniformly attends to the remaining patches. This function significantly

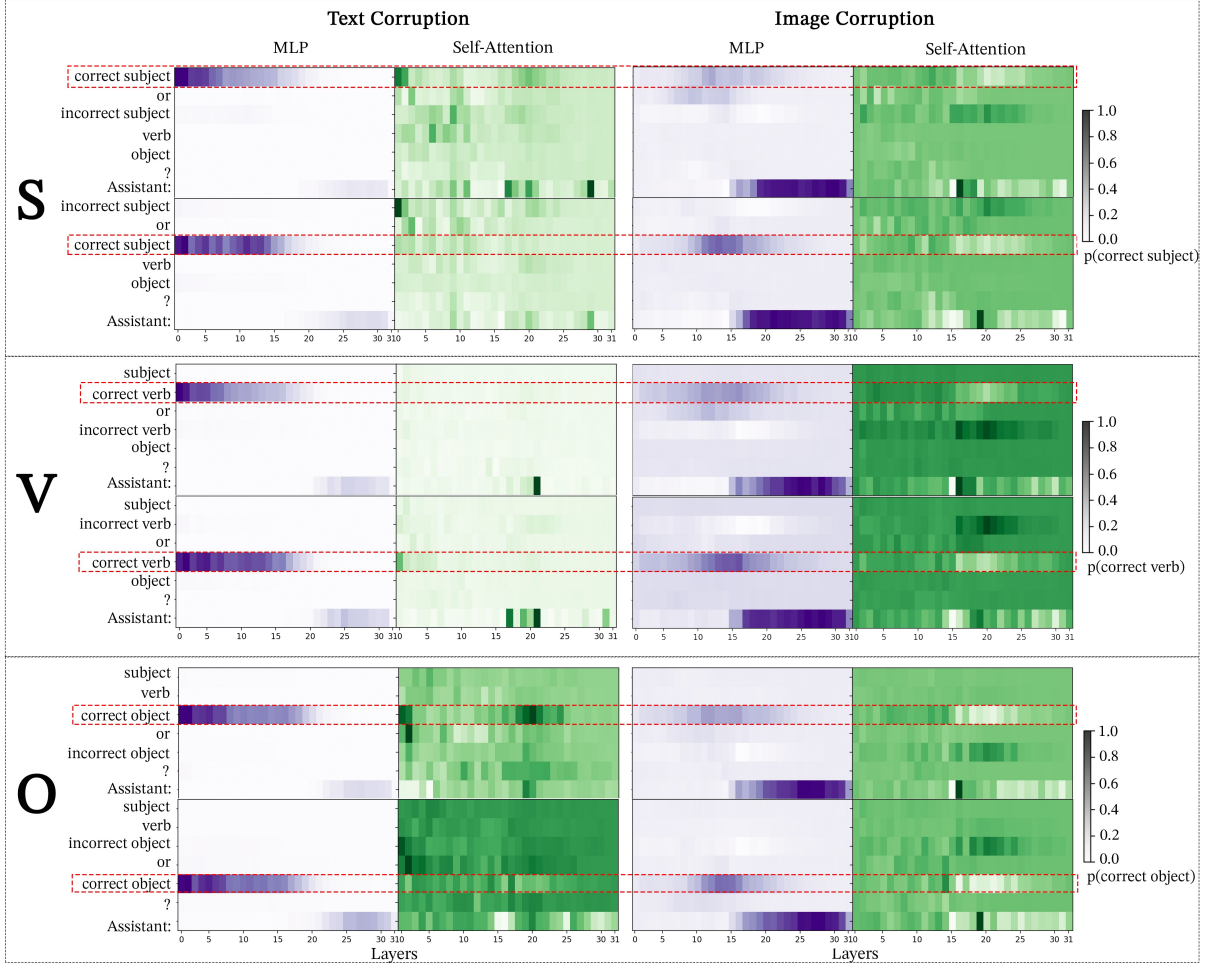


Figure 12: Module-wise activation patching results on LLaVA for the SVO-Probes dataset.

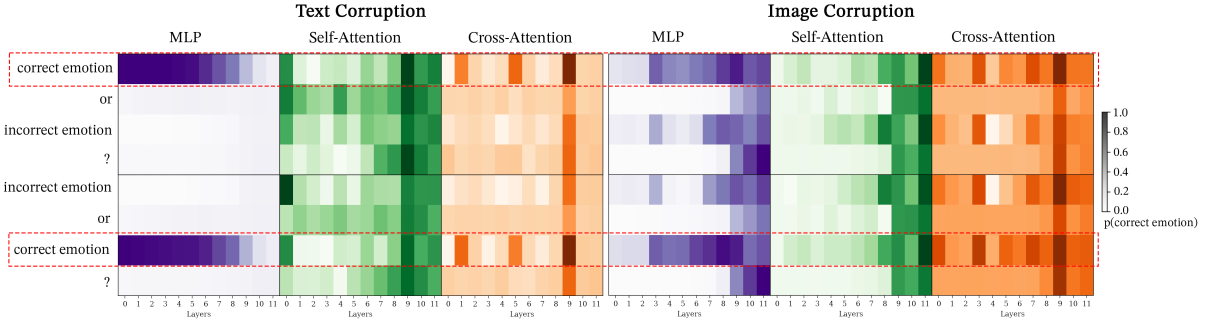


Figure 13: Module-wise activation patching results on BLIP for the Facial Expressions dataset.

reduces the mean of the average cross-attention outputs, thereby enhancing the model’s robustness by filtering out irrelevant visual information. L0.H11 does not produce a significant patching effect when images are corrupted, highlighting its role as a text-only universal cross-attention head.

L3.H0 - Object detection. Cross-attention layer 3, head 0 (L3.H0) performs implicit object detection by associating text tokens with relevant image patches. It consistently demonstrates its im-

portance across all tasks, reliably associating the correct emotion text token with the mouth region in the Facial Expressions dataset, and associating physical objects in images with their corresponding nouns in SVO-Probes and MIT-States. This head’s ability to map input text tokens to associated image patches underscores its critical role in fusing vision-language information.

L5.H3 - object suppression and outlier suppression. Cross-attention head L5.H3 performs ob-

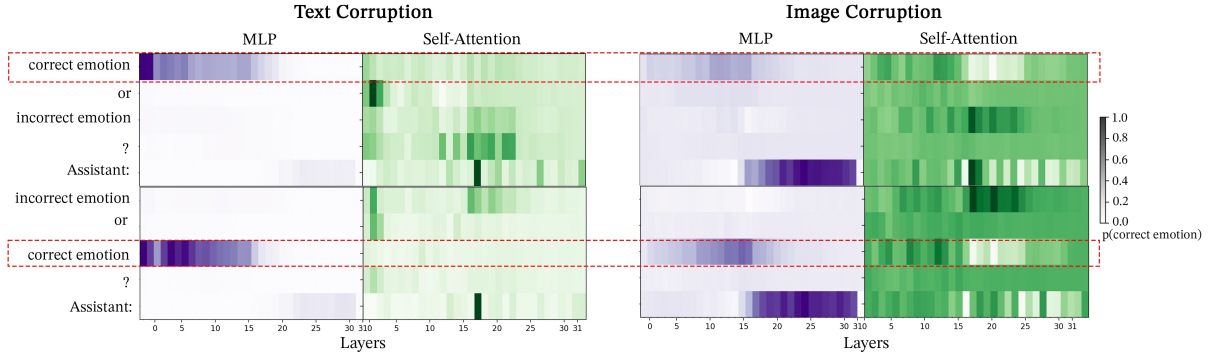


Figure 14: Module-wise activation patching results on LLaVA for the Facial Expressions dataset.

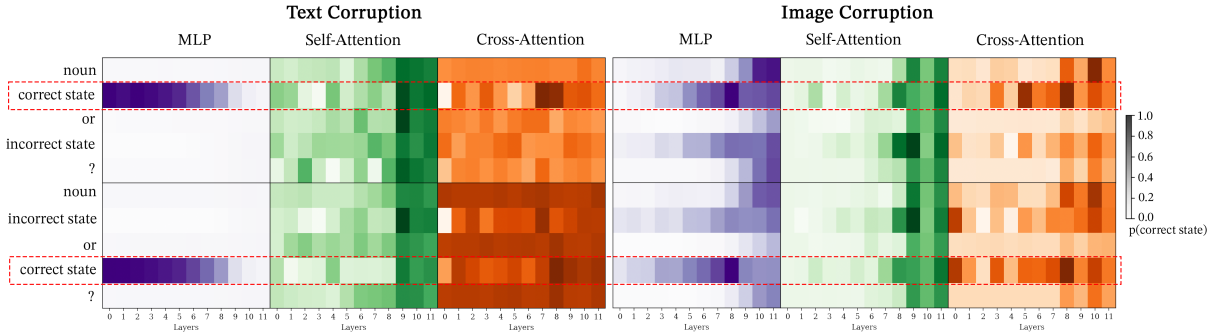


Figure 15: Module-wise activation patching results on BLIP for the MIT States dataset.

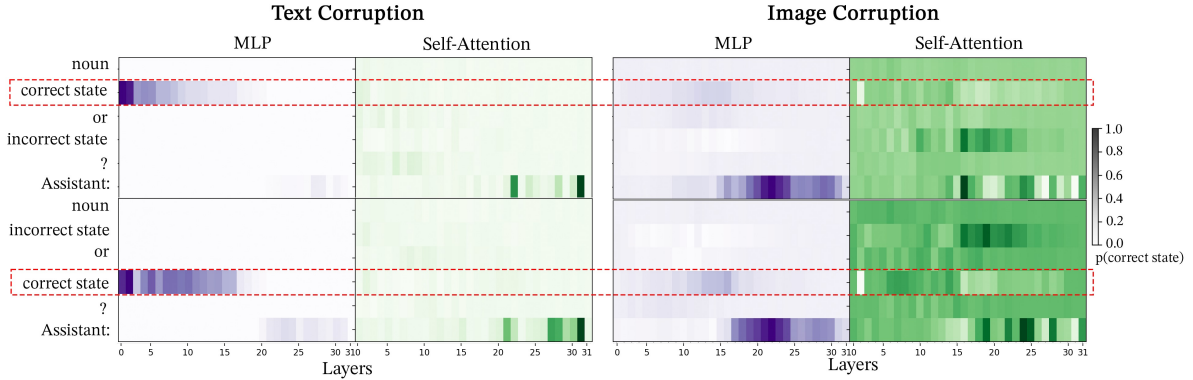


Figure 16: Module-wise activation patching results on LLaVA for the MIT States dataset.

ject suppression by attending away from objects associated with an input token and attending strongly to outlier patches. In SVO-Probes, it attends away from subjects or objects, while in MIT-States, it attends away from relevant portions forming the noun that is changing states. In the Facial Expressions dataset, L5.H3 performs outlier suppression instead of object suppression.

These examples underscore the diverse functions of cross-attention heads within BLIP, each playing a pivotal role in processing and integrating multimodal information. The identification and categorization of these heads provide a deeper understanding of the internal mechanisms driving VLMs,

highlighting the significance of targeted model improvements and adaptations.

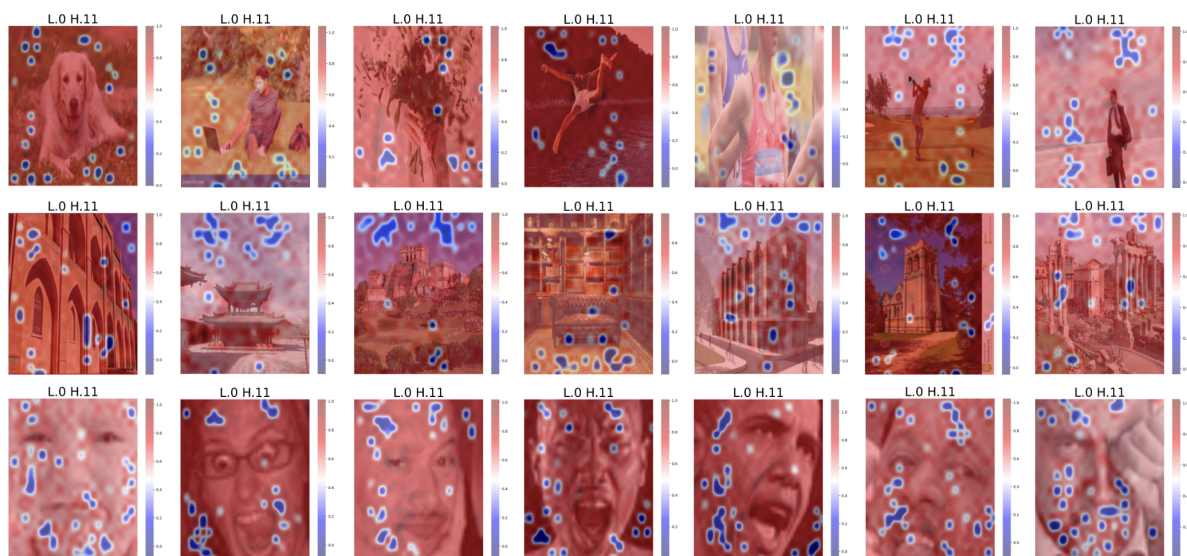
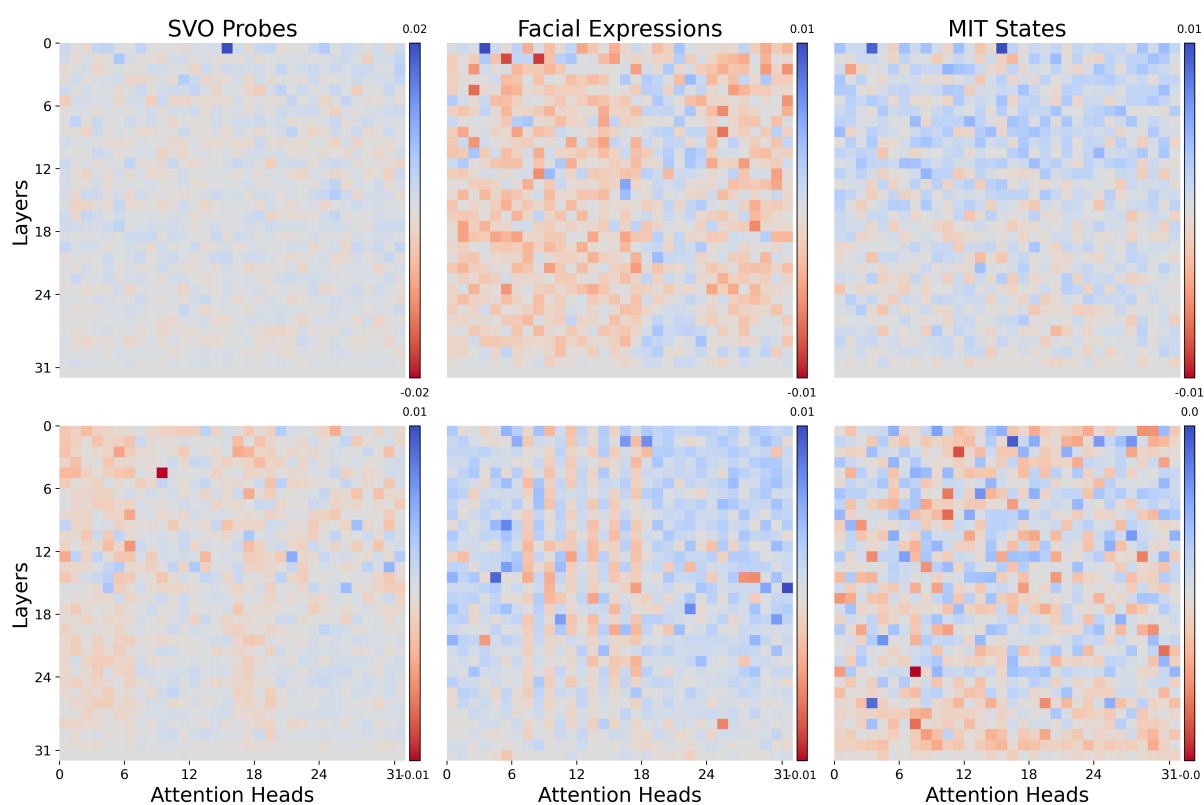


Figure 18: Cross-Attention patterns for universal cross-attention head L0.H11 on a sample of images from each dataset.

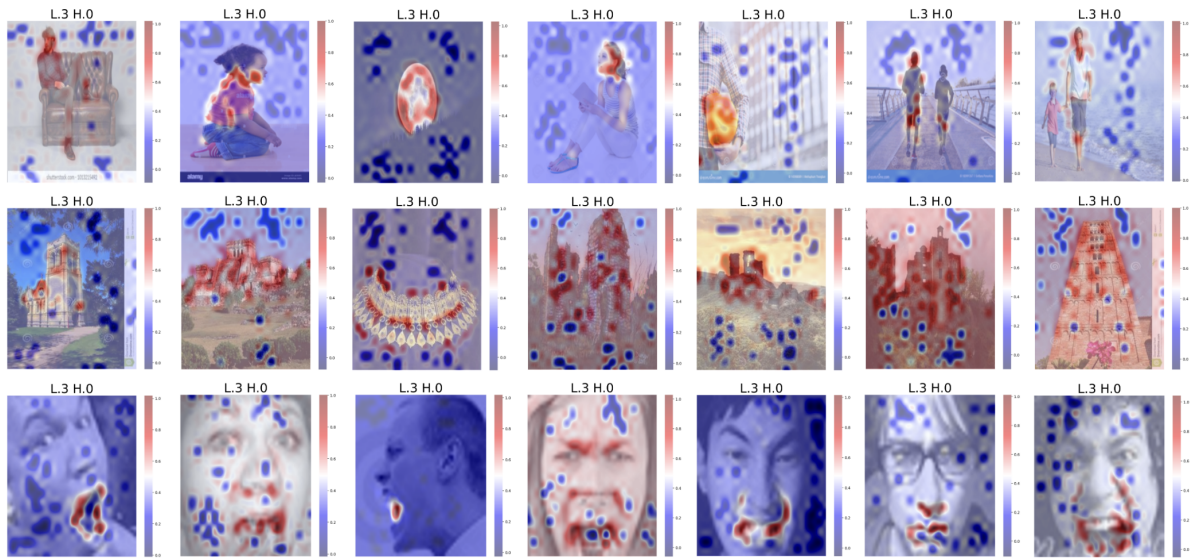


Figure 19: Cross-Attention patterns for universal cross-attention head L3.H0 on a sample of images from each dataset.

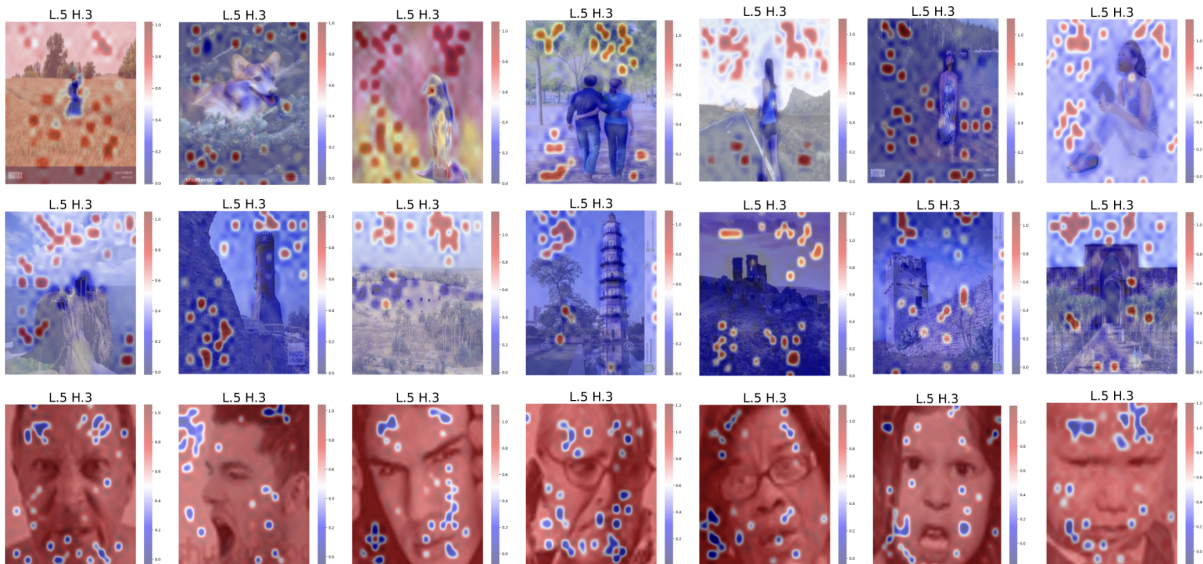


Figure 20: Cross-Attention patterns for universal cross-attention head L5.H3 on a sample of images from each dataset.