Disentangling Hallucinations: Orthogonal Semantic Projection for Robust Interpretability

Emirhan Bilgiç[1,2,3], Baptiste Caramiaux[2], Zhi Yan[1], and Gianni Franchi[4]

[1] U2IS, ENSTA, Institut Polytechnique de Paris, Palaiseau, France
[2] ISIR, Université Sorbonne, Pierre et Marie Curie, Paris, France
[3] Imperial College London, United Kingdom
[4] AMIAD, Pôle Recherche, Palaiseau, France

e.bilgic26@imperial.ac.uk

Abstract. As Vision-Language Models are increasingly deployed in safety-critical applications, the trustworthiness of their explanations becomes crucial. Explainable AI (XAI) methods for Vision-Language Models often suffer from *explanation hallucination*, where attribution maps highlight prominent image regions even when prompted with incorrect text descriptions (e.g., highlighting a dog when prompted "cat"). Although this problem is widespread, a formal mathematical analysis of XAI methods and CLIP embeddings is largely missing in the literature. We demonstrate that this phenomenon is not specific to a single architecture but is a fundamental consequence of Linear Semantic Leakage in high-dimensional embedding spaces. We propose a unified theoretical framework, Linear Semantic Attribution (LSA), which generalizes across discriminative methods. We introduce Orthogonal Semantic Projection (OSP), a geometric intervention that utilizes the residual property of OMP to disentangle unique semantic signals from shared concepts. We prove theoretically and demonstrate empirically that OSP minimizes hallucination by orthogonalizing the query vector against distractor concepts, rendering the attribution model blind to shared features while preserving fidelity for correct prompts. The project website, including the code and animations, is available at: `https://emirhanbilgic.github.io/Orthogonal-Semantic-Projection`

Keywords: Vision-Language Models, Explainable AI, Hallucination, Orthogonal Matching Pursuit, CLIP, SigLIP, Diffusion Models

1 Introduction

Vision-Language Models (VLMs) and multimodal diffusion models have become central to modern computer vision. Well-known examples include CLIP [33] and SigLIP [39] for image–text understanding, as well as generative architectures such as Stable Diffusion [35], Flux [1], and PixArt-α [5]. These systems have enabled significant progress in areas such as autonomous driving [40], medical image synthesis [17], and zero-shot robotics [26]. However, as they move from

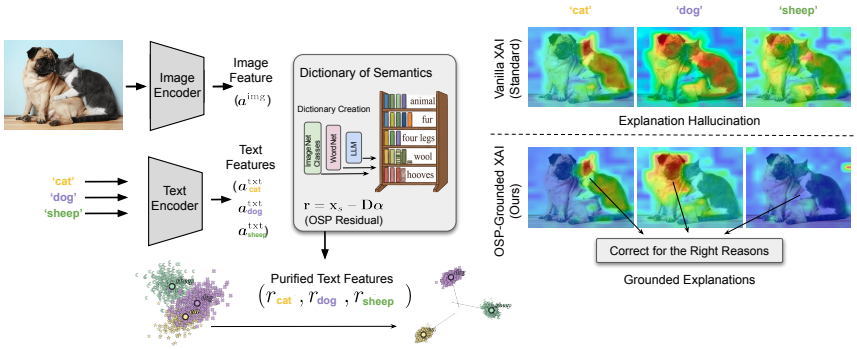

Fig. 1: Overall pipeline of OSP. Our method performs a geometric intervention in the multimodal embedding space to disentangle shared semantic components that cause hallucination, resulting in more faithful and grounded explanations.

research prototypes to real-world deployment, a critical challenge emerges: they are difficult to interpret. Their internal representations function largely as "black boxes," undermining trust, especially in safety-critical environments.

To better understand these models, researchers commonly apply post-hoc explainability methods originally developed for standard Deep Neural Networks (DNNs). These methods produce saliency maps or attribution masks that highlight important regions in an image. Yet when applied to multimodal models, a new and severe problem appears: *explanation hallucination*. Saliency methods often highlight objects that are completely absent from the image, for instance, highlighting a dog when prompted with "cat." This undermines the trustworthiness of Explainable AI (XAI) and, more broadly, reduces the possibility of effectively debugging AI models and makes AI less safe.

Hallucination is commonly defined as generated content that is incorrect or not grounded in the input [14]. In this work, we show that hallucination does not only affect model *outputs*: it also affects *explanations*. We define **explanation hallucination** as a situation where a saliency map highlights image regions that are not truly related to the given text prompt, as illustrated in Fig. 1, where the explanation hallucinates the presence of sheep. This problem is critical because the explanation appears convincing but is not correctly grounded in the text: the model may seem correct, but it is "right for the wrong reasons." We stress that here "related" denotes *semantic faithfulness* rather than spatial *localization*: a saliency map can be sharply localized on a salient object yet remain unfaithful when that object does not match the prompt. We regard an explanation as faithful only if the highlighted evidence is specific to the queried concept. This distinction is measurable: because the CLIP/SigLIP zero-shot score for an *absent* class is small, a faithful map should assign that class little saliency, whereas explanation hallucination assigns high saliency to a concept the model itself

scores as absent. Explanation hallucination is therefore a failure of faithfulness, distinct from poor localization.

We argue that this issue stems from the geometry of the multimodal embedding space. Image and text features are represented in a shared high-dimensional space. Because these embeddings are not orthogonal, semantically related but distinct concepts share common directional components (see Fig. 2). When saliency maps are computed using these embeddings, these shared components cause the method to highlight irrelevant regions, and especially "wrong objects": producing hallucinated explanations. While not addressing the hallucination of attribution maps, there have been multiple works on image embeddings [9,28,29], but to the best of our knowledge, we are the first to provide such work on text embeddings to improve attribution XAI methods.

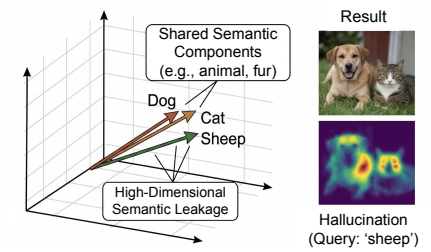

Fig. 2: 3D text embeddings and hallucination. Without disentanglement, the distractor is "distracting" the heatmaps.

In this paper, we introduce **Orthogonal Semantic Projection (OSP)**, a geometric intervention designed to reduce hallucinations in saliency-based explanations. OSP is grounded in a simple geometric idea: before computing the saliency map, we remove the shared semantic components in the text embedding that cause hallucination.

To achieve this, we use a dictionary learning framework based on Orthogonal Matching Pursuit (OMP) [31]. This sparse decomposition allows us to isolate and remove the parts of the embedding that introduce semantic leakage. We then compute the saliency map using a refined, orthogonalized representation of the text query. Importantly, OSP acts as a **generalized framework** and is *plug-and-play*: it can be applied to a wide range of Vision-Language models and saliency methods without retraining the underlying model.

Crucially, this represents a fundamentally new paradigm for interpretability: instead of decomposing the image embedding space [10] to find visual concepts, we mathematically decompose the *text* embedding space to isolate pure semantic directions. We summarize our contributions as follows:

1. **Geometric Analysis of Hallucination:** We provide a formal mathematical framework that bridges the gap between grounding failures and hallucinatory outputs, characterizing hallucination as a function of the latent misalignment (i.e. non-orthogonality) between semantic embedding directions and offering a geometric interpretation of how grounding errors propagate into nonsensical attributions.

2. **A Universal XAI Technique:** OSP is compatible with multiple architectures, including CLIP, SigLIP, and diffusion-based models, and can be integrated with different saliency methods as a plug-and-play module.
3. **Extensive Validation:** We evaluate OSP across 3 foundational models and 5 state-of-the-art XAI methods, demonstrating consistent improvements.
4. **Practical Use Cases and User Study:** We demonstrate improved interpretability reliability, backed by empirical results and a user study.

2 Related Work

2.1 Vision-Language Encoder Attribution Methods

Attribution methods and saliency maps aim to demystify the "black box" nature of Vision-Language Models (VLMs) by grounding textual concepts in visual regions. Discriminative approaches adapt gradient-based strategies to multimodal architectures. For instance, GradCAM [36] has been successfully applied to CLIP to identify visual features that maximize similarity with specific text queries. For Vision Transformers (ViTs), methods like LeGrad [2] provide feature formation sensitivity by tracing gradients to internal patch embeddings, while CheferCAM [4] directly propagates relevance scores using attention maps and their gradients. Generative models, such as Stable Diffusion, rely on methods like DAAM [38], which aggregates cross-attention maps to spatially localize the influence of prompt words. Recent works such as TextSpan [10] evaluate concept attribution by decomposing image representations into text-based constituents. However, unlike these existing explainability approaches that accept raw entangled query embeddings and refine spatial attribution post-hoc or within bottlenecks, our method performs a structured geometric intervention directly in the continuous representation space prior to attribution to proactively resolve spatial leakage caused by shared semantic similarities.

2.2 Hallucination in Vision-Language Models

Despite their remarkable zero-shot capabilities, VLMs and Large Vision-Language Models (LVLMs) frequently suffer from different types of hallucinations. The severity of this issue has motivated the creation of targeted evaluation benchmarks. Several works tried to address object hallucination in the Visual Question Answering (VQA) setting: the CHAIR metric [34] quantifies object hallucinations in image captioning, the POPE benchmark [21] evaluates object hallucination in LVLMs through yes-no questions, and THRONE [16] benchmarks hallucination in free-form generations. To mitigate concept hallucination within Concept Bottleneck Models (CBMs), Kazmierczak et al. [18] introduce CHILI, a technique that disentangles image embeddings to localize concept pixels. However, other benchmarks focus on VQA, and CHILI focuses on concept prediction within CBMs and does not address the hallucination of attribution maps. Our method specifically targets *object attribution hallucination*. While these diagnostic tools highlight the prevalence of hallucination as an empirical phenomenon to

be benchmarked or mitigated via targeted disentanglement, our work uniquely establishes a formal, mathematical link connecting the root cause of these errors backward into the linear geometry of the multimodal representation space itself, while it can be easily integrated with existing attribution methods.

2.3 Representation Geometry of Vision-Language Encoders

Emerging research characterizes the high-dimensional latent spaces of networks like CLIP not as uniform structures, but as complex geometric manifolds. The Linear Representation Hypothesis [30] suggests that high-level concepts and hierarchical structures are encoded as linear directions. Furthermore, the Platonic Representation Hypothesis [15] posits that diverse architectures naturally converge toward a shared geometric model of reality. Within contrastive models like CLIP, geometric phenomena such as the "modality gap" [22] reveal that image and text embeddings reside on separated, offset hyperspheres, severely restricting the angular portion of the space utilized by the model. To disentangle these compressed spaces, Sparse Autoencoders (SAEs) [9] have recently been employed to extract monosemantic, interpretable directions from dense activations. While prior works focus on analyzing structures like the modality gap or sparse features, OSP uniquely utilizes these geometric insights as a mitigation tool, strictly dynamically orthogonalizing directions at inference time to correct flawed model behavior and concept bleeding.

3 The Geometry of Hallucination

3.1 Background on Saliency Map Methods for Multimodal Models

Notations. We consider a DNN denoted by $f_{\boldsymbol{\omega}}(\cdot)$, where $\boldsymbol{\omega}$ represents the model parameters. In this paper, we focus on multimodal encoders [33, 39] and text-conditioned diffusion models [35] that process two inputs: an image $\mathbf{x}^{\text{img}}$ and a text prompt $\mathbf{x}^{\text{txt}}$. We do not consider autoregressive multimodal large language models (MLLMs). In diffusion-based models, the image input initially consists of noise that is progressively refined during the generation process.

Feature Extractors and Activations. We define the saliency maps based on two distinct categories of model activations: **(1) Forward Activations:** Let $\mathbf{A}_{l,h}^{\text{MHA}}(\mathbf{x}^{\text{img}})$ denote the activation map corresponding to the h-th attention head in the l-th layer of the Transformer. Similarly, let $\mathbf{A}_{l}^{\text{MLP}}(\mathbf{x})$ represent the output activations of the Multi-Layer Perceptron (MLP) sub-layer at layer l. We denote the complete set of forward activations as $\{\mathbf{A}_{l,h}^{\text{MHA}}(\mathbf{x}^{\text{img}}), \mathbf{A}_{l}^{\text{MLP}}(\mathbf{x}^{\text{img}})\}_{l,h=1}^{L,H}$. **(2) Backward Activations:** To incorporate local sensitivity and importance weighting, we utilize gradient-based signals derived from the backward pass. Let $S_c = \langle \mathbf{a}_c^{\text{txt}}, \mathbf{a}^{\text{img}} \rangle$ be the cosine similarity score for class or concept c. Following the formulation in LeGrad [2] and CheferCAM [4], we define the backward activation set as $\{\nabla_{\mathbf{A}_{l,h}^{\text{MHA}}} S_c\}_{l,h=1}^{L,H}$, where $\nabla_{\mathbf{A}_{l,h}^{\text{MHA}}} S_c$ represents the gradient of the class similarity score with respect to the h-th attention head in the l-th layer.

Text and Image Embeddings. Let us consider a set of textual classes or prompts: $\{\mathbf{x}_c^{\text{txt}}\}_c$. Each text input is mapped to an embedding: $\{\mathbf{a}_c^{\text{txt}}\}_c$, $\mathbf{a}_c^{\text{txt}} \in \mathbb{R}^h$, where h is the embedding dimension. Similarly, we denote by $\mathbf{a}_j^{\text{img}}(\mathbf{x}^{\text{img}})$ the image embedding extracted at layer j. Depending on the architecture, for example, for CLIP or SigLIP, this may correspond to the class token, and for diffusion models, this may correspond to an averaged image token or latent representation. For simplicity, we write $\mathbf{a}_j^{\text{img}}$.

General Formulation of Saliency Maps. Most saliency methods aim to produce a single spatial importance map $\mathbf{V}_c(\mathbf{x})$ for a given class or prompt c. We observe that many methods can be written as a weighted combination of activation maps, where the weights depend on the text embedding or on its interaction with the image embedding. Formally, the saliency map can be expressed as:

$$\mathbf{V}_c(\mathbf{x}^{\text{img}}) = \begin{cases} \mathbf{w}^c\, \mathbf{A}_L^{\text{MLP}}(\mathbf{x}^{\text{img}}) \text{ with } \mathbf{w}^c = \mathbf{a}_c^{\text{txt}}, & (\textbf{CAM}), \\ \mathbf{w}^c\, \mathbf{A}_L^{\text{MLP}}(\mathbf{x}^{\text{img}}) \text{ with } \mathbf{w}^c = \frac{\partial}{\partial \mathbf{a}_c^{\text{txt}}} \langle \mathbf{a}_c^{\text{txt}}, \mathbf{a}_d^{\text{img}} \rangle, & (\textbf{GradCAM}\ [36]), \\ \sum_{l,h} \mathbf{w}_j\, \nabla_{\mathbf{A}_{l,h}^{\text{MHA}}} S_c \text{ with } \mathbf{w}_j = 1/(HL), & (\textbf{LeGrad}\ [2]), \\ \left[\prod_{l=1}^{L} \left(\mathbf{I} + \frac{1}{H} \sum_h \left(\nabla_{\mathbf{A}_{l,h}^{\text{MHA}}} S_c \odot \mathbf{A}_{l,h}^{\text{MHA}}(\mathbf{x}^{\text{img}}) \right)^+ \right) \right]_{\text{CLS}}, & (\textbf{CheferCAM}\ [4]), \\ \sum_h \mathbf{w}_j \left(\nabla_{\mathbf{A}_{L,h}^{\text{MHA}}} S_c \odot \mathbf{A}_{L,h}^{\text{MHA}}(\mathbf{x}^{\text{img}}) \right)^+, \text{ with } \mathbf{w}_j = 1/H, & (\textbf{AttentionCAM}\ [3]), \\ \frac{1}{T} \sum_{t=1}^{T} \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} \mathbf{A}_{(t,h)}^{\text{cross}}(\mathbf{x}^{\text{img}}, c), & (\textbf{DAAM}\ [38]), \end{cases} \tag{1}$$

where: $(\cdot)^+$ denotes the ReLU activation function, $\odot$ denotes the element-wise Hadamard product, $\mathbf{A}_{(t,h)}^{\text{cross}}(\mathbf{x}, c)$ refers to the cross-attention map for token c at diffusion timestep t and head h. This formulation shows that most multimodal saliency methods follow a similar structure where, before the aggregation, the multimodal representation can be viewed as living in a space of dimension $d \times H \times L$ (where d is the embedding dimension), and then in a space of 1 dimension. Therefore, saliency extraction can be interpreted as a dimensionality reduction operation that maps this high-dimensional multimodal representation to a single scalar importance value per spatial location. This unified view will allow us to analyze how interactions between text and image embeddings can introduce semantic leakage and lead to hallucinated explanations.

3.2 From Linear Attribution to Semantic Leakage

Linear Dependence on Text Embeddings. Most saliency methods for vision-language models can be written under the general form

$$\mathbf{V}_c(\mathbf{x}) = f\big(\bar{\mathbf{a}}^{\text{img}}, \mathbf{a}_c^{\text{txt}}\big), \tag{2}$$

where the operator $f(\cdot, \cdot)$ characterizes the interaction mechanism between the visual features and the text query, mapping the joint multimodal representation into the spatial saliency domain as described in Eq. (1). In many practical instantiations (e.g., similarity-based scores or their gradients), this interaction

behaves linearly with respect to the text embedding, or can be locally approximated as such (see Appendix A.1). We rely on this assumption for the rest of this section.

Geometric Structure of the Embedding Space. Consider two classes A and B with text embeddings $\mathbf{a}_A^{\text{txt}}, \mathbf{a}_B^{\text{txt}} \in \mathbb{R}^d$. In contrastive models, embeddings are normalized:

$$\|\mathbf{a}_A^{\text{txt}}\|_2 = \|\mathbf{a}_B^{\text{txt}}\|_2 = 1. \tag{3}$$

Let θ_{AB} denote the angle between them:

$$\langle \mathbf{a}_A^{\text{txt}}, \mathbf{a}_B^{\text{txt}} \rangle = \cos(\theta_{AB}). \tag{4}$$

Since semantically related concepts share directions in embedding space, $\cos(\theta_{AB})$ is typically non-zero. To empirically evaluate this, we computed the pairwise cosine similarities between the text embeddings of all 1,000 ImageNet classes. Across all 499,500 unique pairs, the similarity is never zero: the minimum similarity is 0.0967, with a mean of 0.5925, a median of 0.6082, and a maximum of 1.0000. This confirms that orthogonal concept embeddings are practically non-existent in standard VLM representation spaces. We denote by $\mathbf{a}_{B\perp A}^{\text{txt}}$ the component of $\mathbf{a}_B^{\text{txt}}$ orthogonal to $\mathbf{a}_A^{\text{txt}}$.

Definition (Hallucinated Attribution). Let image $\mathbf{x}$ belong to class A. Let $\mathbf{V}_c(\mathbf{x})$ denote the saliency map extracted for class c. We say that hallucination occurs if

$$mean_{\mathbf{x}}\mathbf{V}_B(\mathbf{x}) \;\geq\; mean_{\mathbf{x}}\mathbf{V}_A(\mathbf{x}), \tag{5}$$

even though the image contains only class A.

Theorem 1 (Linear Semantic Leakage). *Let image* $\mathbf{x}$ *contain an object of class* A*. Let us denote* $\mathbf{a}^{img}$ *its image embedding. If*

$$\langle \mathbf{a}_A^{txt}, \mathbf{a}_B^{txt} \rangle \neq 0, \tag{6}$$

then the saliency map $\mathbf{V}_B(\mathbf{x})$ *contains a component proportional to* $\mathbf{V}_A(\mathbf{x})$*.*

Proof. Because $\|\mathbf{a}_A^{\text{txt}}\|_2 = 1$, we decompose $\mathbf{a}_B^{\text{txt}}$ via orthogonal projection:

$$\mathbf{a}_B^{\text{txt}} = \langle \mathbf{a}_B^{\text{txt}}, \mathbf{a}_A^{\text{txt}} \rangle \mathbf{a}_A^{\text{txt}} + \mathbf{a}_{B\perp A}^{\text{txt}}. \tag{7}$$

Using the cosine identity:

$$\mathbf{a}_B^{\text{txt}} = \cos(\theta_{AB})\, \mathbf{a}_A^{\text{txt}} + \mathbf{a}_{B\perp A}^{\text{txt}}. \tag{8}$$

All attribution methods considered in Eq. (1) are linear in the text embedding and can be written as

$$\mathbf{V}_c(\mathbf{x}) = f\left(\bar{\mathbf{a}}^{\text{img}}, \mathbf{a}_c^{\text{txt}}\right). \tag{9}$$

where $\bar{\mathbf{a}}^{\text{img}}$ denotes the specific image representation (e.g., intermediate activations or normalized embeddings) utilized by the respective algorithm. The

operator $f(\cdot,\cdot)$ characterizes the interaction mechanism between the visual features and the text query. We assume that $f(\cdot,\cdot)$ is linear with the text input, hence we have:

$$\mathbf{V}_B(\mathbf{x}) = f\left(\bar{\mathbf{a}}^{\text{img}}, \left(\cos(\theta_{AB})\mathbf{a}_A^{\text{txt}} + \mathbf{a}_{B\perp A}^{\text{txt}}\right)\right). \tag{10}$$

By linearity:

$$\mathbf{V}_B(\mathbf{x}) = \cos(\theta_{AB})\mathbf{V}_A(\mathbf{x}) + f\left(\bar{\mathbf{a}}^{\text{img}}, \mathbf{a}_{B\perp A}^{\text{txt}}\right). \tag{11}$$

The first term is proportional to the true saliency map $\mathbf{V}_A(\mathbf{x})$. This term is the *ghost signal*.

Implications. This result shows that hallucination is directly controlled by $\cos(\theta_{AB})$. We empirically validate it in Appendix B: "Correlation between Cosine Similarity and Hallucination". If the angle between embeddings is small, the ghost signal can dominate the residual term. Therefore, reducing hallucination amounts to reducing the projection of a query embedding onto unwanted semantic directions. This phenomenon demonstrates the semantic contamination of class/concept A's saliency map by the representation of class/concept B. Such interference suggests that the attribution mechanism fails to isolate class/concept-specific features when embeddings are non-orthogonal. We provide a comprehensive extension of this analysis to multi-class scenarios in Appendix A.2.

4 Methodology

4.1 Sparse Dictionary Decomposition of Text Embeddings

We formulate semantic purification as a sparse decomposition problem in the joint embedding space. Let $\mathbf{a}_{\text{target}}^{\text{txt}} \in \mathbb{R}^d$ denote the unit-normalized embedding of a target concept obtained from a pretrained vision-language model. Let $\mathbf{D} = [\mathbf{a}_{D_1}^{\text{txt}}, \ldots, \mathbf{a}_{D_k}^{\text{txt}}] \in \mathbb{R}^{d\times k}$ be a dictionary of semantics embeddings, each normalized to unit norm.

In sparse representation theory [7], a signal $\mathbf{x}_s \in \mathbb{R}^d$ can be approximated as

$$\mathbf{x}_s \approx \mathbf{D}\boldsymbol{\alpha}, \tag{12}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^k$ is sparse. The residual

$$\mathbf{r} = \mathbf{x}_s - \mathbf{D}\boldsymbol{\alpha} \tag{13}$$

captures the component of the signal that cannot be explained by the selected atoms.

In our setting, we treat $\mathbf{a}_{\text{target}}^{\text{txt}}$ as the signal and the distractor embeddings as dictionary atoms. Our objective is not reconstruction accuracy per se, but rather *orthogonal purification*: we aim to remove the semantic components shared with distractors and retain only the unique semantic direction of the target.

4.2 Orthogonal Matching Pursuit for Semantic Purification

To compute this decomposition we use Orthogonal Matching Pursuit (OMP) [31], a greedy algorithm that iteratively selects atoms most correlated with the current residual.

Starting from $\mathbf{r}^{(0)} = \mathbf{a}^{\text{txt}}_{\text{target}}$, OMP performs the following steps:

1. Select the atom most correlated with the current residual:

$$j^\star = \arg\max_{j \notin \Lambda} \left| \langle \mathbf{r}^{(t-1)}, \mathbf{a}^{\text{txt}}_{D_j} \rangle \right|. \tag{14}$$

2. Update the active set Λ of selected atoms ($\mathbf{D}_\Lambda = [\mathbf{a}^{\text{txt}}_{D_j}]_{j \in \Lambda^{(t)}}$).
3. Recompute the residual via orthogonal projection:

$$\mathbf{r}^{(t)} = \mathbf{a}^{\text{txt}}_{\text{target}} - \mathbf{D}_\Lambda (\mathbf{D}_\Lambda^\top \mathbf{D}_\Lambda)^{-1} \mathbf{D}_\Lambda^\top \mathbf{a}^{\text{txt}}_{\text{target}}. \tag{15}$$

A fundamental property of OMP is:

$$\langle \mathbf{r}^{(T)}, \mathbf{a}^{\text{txt}}_{D_j} \rangle = 0, \qquad \forall j \in \Lambda, \tag{16}$$

meaning the final residual is strictly orthogonal to all selected distractors.

We then normalize:

$$\mathbf{r} = \frac{\mathbf{r}^{(T)}}{\|\mathbf{r}^{(T)}\|_2}. \tag{17}$$

4.3 Orthogonal Semantic Projection (OSP)

We define **Orthogonal Semantic Projection (OSP)** as the use of the OMP residual as the purified query embedding.

Definition. Given a target embedding $\mathbf{a}^{\text{txt}}_{\text{target}}$ and a dictionary of semantics $\mathbf{D}$, OSP outputs

$$\mathbf{r} = \text{OMP-residual}(\mathbf{a}^{\text{txt}}_{\text{target}}, \mathbf{D}), \tag{18}$$

which satisfies

$$\langle \mathbf{r}, \mathbf{a}^{\text{txt}}_{D_i} \rangle = 0, \quad \forall i \in \Lambda. \tag{19}$$

Intuitively, OSP decomposes the target embedding into

$$\mathbf{a}^{\text{txt}}_{\text{target}} = \underbrace{\mathbf{D}\boldsymbol{\alpha}}_{\text{shared semantics}} + \underbrace{\mathbf{r}}_{\text{unique semantics}}, \tag{20}$$

and retains only the residual component.

Because multimodal saliency methods are linear in the text embedding, replacing $\mathbf{a}^{\text{txt}}_{\text{target}}$ by $\mathbf{r}$ eliminates ghost signals originating from shared semantic directions.

4.4 Application to Attribution Methods

Discriminative Methods. For CAM, GradCAM, LeGrad, CheferCAM, and AttentionCAM, the saliency map can be written as $\mathbf{V}_c(\mathbf{x}) = f\left(\bar{\mathbf{a}}^{\text{img}}, \mathbf{a}_c^{\text{txt}}\right)$. Replacing $\mathbf{a}_c^{\text{txt}}$ with $\mathbf{r}$ yields

$$\mathbf{V}_c^{\text{OSP}}(\mathbf{x}) = f\left(\bar{\mathbf{a}}^{\text{img}}, \mathbf{r}\right) \tag{21}$$

If a distractor concept D_i explains a region of the image, then $\langle \mathbf{r}, \mathbf{a}_{D_i}^{\text{txt}} \rangle = 0$ implies that the corresponding contribution vanishes. Thus, ghost saliency components are removed by construction.

Diffusion-Based Attribution (DAAM). In diffusion models, attribution arises from cross-attention maps. Let $\mathbf{K}_{\text{target}}^{(l,h)}$ denote the key vector associated with the target token in layer l, head h. We apply OSP in key space:

$$\mathbf{r}^{(l,h)} = \mathbf{K}_{\text{target}}^{(l,h)} - \beta \sum_i \text{proj}_{\hat{\mathbf{K}}_{D_i}^{(l,h)}}\left(\mathbf{K}_{\text{target}}^{(l,h)}\right), \qquad \hat{\mathbf{K}}_{D_i}^{(l,h)} = \frac{\mathbf{K}_{D_i}^{(l,h)}}{\|\mathbf{K}_{D_i}^{(l,h)}\|}, \tag{22}$$

where the projections onto the normalized distractor keys $\hat{\mathbf{K}}_{D_i}^{(l,h)}$ are removed successively and β controls the suppression strength.

The purified keys are substituted before attention softmax. Since attention scores depend linearly on keys before normalization, orthogonality ensures that distractor-aligned attention weights are suppressed.

4.5 Dictionary of Semantics Construction

For each target concept, we extract the *dictionary of semantics*: a structured collection of semantically related distractor concepts whose text embeddings form the atoms of the dictionary of semantics $\mathbf{D}$.

To build the dictionary of semantics $\mathbf{D}$, we explored four accessible dictionary creation strategies: (1) using ImageNet classes, (2) incorporating ImageNet classes with WordNet semantic relations, and generating customized dictionaries via accessible Large Language Models, specifically (3) Gemini 3 Flash and (4) GPT-OSS 120B. Detailed definitions for these strategies are provided in Appendix C. For all approaches, the hyperparameters kept identical for both positive and negative cases. Furthermore, we filter out any atoms that exhibit excessively high cosine similarity to the target concept to avoid deleting the core semantic information. The hyperparameters for different dictionaries vary in a small space. More can be found in Appendix D.

As shown in Table 2, all dictionaries consistently perform well and improve the AUROC. The dictionary generated with Gemini 3 Flash is selected as our primary approach. We carefully designed the LLM prompt to generate dictionary atoms structured into three specific semantic groups:

- **Group 1: Visual Confusers** (15 concepts) – objects that share strong visual features with the target concept.
- **Group 2: Co-occurring Context** (15 concepts) – objects or background elements frequently found in the same scenes as the target.
- **Group 3: Semantic Hierarchy** (10 concepts) – structurally related concepts comprising 5 hypernyms and 5 hyponyms.

Our results are fully reproducible: the generated dictionary is open source and publicly available, and the exact prompt used to recreate the dictionary is provided in Appendix H.

5 Experiments

5.1 Experimental Setups

Overall, we test our framework across 3 different models (CLIP [33], SigLIP [39], Stable Diffusion 2 [35]) and 5 different methods (GradCAM [36], LeGrad [2], CheferCAM [4], AttentionCAM [3], DAAM [38]). We evaluate the empirical performance of these methods on the ImageNet-Segmentation [11], a curated benchmark containing 4,276 images from the ImageNet validation set equipped with precise pixel-level annotations. We also conduct extra experiments on MS COCO, and the results are available in Appendix D. To further validate our approach, we conducted a human study using a dataset of 1.5k heatmaps. For this evaluation, we use the five methods, evaluated both with and without their corresponding OSP heatmaps across three different concepts/classes. We use 50 randomly sampled images from the validation sets of Pascal VOC 2012 [8] and MS COCO [23], making it $5 \times 2 \times 3 \times 50 = 1{,}500$ heatmaps. We specifically selected images that contain at least 2 unique objects to effectively test multi-concept disambiguation in clustered scenes. More information about this experiment is provided in Section 6.2 and Appendix F.

5.2 Quantitative Results

We quantitatively evaluate OSP on the ImageNet-Segmentation benchmark using four standard metrics: mean Intersection-over-Union (mIoU), mean Average Precision (mAP), pixel accuracy, and the Area Under the ROC Curve (AUROC). Detailed results are presented in Table 1, where the dictionary of semantics was extracted using Gemini. For each method-model pair, we generate heatmaps under two distinct conditions: **Positive prompts**: The queried concept is present in the image. **Negative prompts**: The queried concept is absent. A faithful attribution method should yield high scores for positive prompts and low scores for negative prompts. OSP is designed to widen this gap by enhancing positive scores while suppressing negative ones. In particular, we utilize AUROC to validate our theoretical claims regarding hallucination. As formalized in Eq. 5, explanation hallucination occurs when the mean saliency of an absent distractor class B is comparable to or exceeds that of a present class A. Since AUROC

Table 1: Detailed mIoU, mAP, and AUROC before and after OSP on ImageNet-Segmentation using the Gemini 3 Flash dictionary. Δ columns show the change introduced by OSP. For positive prompts (↑ desired), OSP should increase scores; for negative prompts (↓ desired), OSP should decrease them. Favorable changes are highlighted in **bold**.

Method	Model	Prompt	mIoU			mAP			AUROC		
			Base	+OSP	Δ	Base	+OSP	Δ	Base	+OSP	Δ
LeGrad	CLIP	Pos ↑	58.66	61.33	**+2.67**	82.49	85.74	**+3.25**	79.62	80.64	**+1.02**
		Neg ↓	40.88	40.74	**-0.14**	67.99	70.86	+2.87			
	SigLIP	Pos ↑	49.51	51.18	**+1.67**	78.32	78.78	**+0.46**	74.42	74.73	**+0.31**
		Neg ↓	37.27	34.68	**-2.59**	63.63	62.84	**-0.79**			
CheferCAM	CLIP	Pos ↑	48.71	51.32	**+2.61**	80.36	82.60	**+2.24**	77.63	80.14	**+2.51**
		Neg ↓	44.88	42.99	**-1.89**	78.74	80.12	+1.38			
	SigLIP	Pos ↑	37.66	39.60	**+1.94**	73.49	75.95	**+2.46**	55.47	62.64	**+7.17**
		Neg ↓	36.65	38.45	+1.80	72.38	74.67	+2.29			
AttentionCAM	CLIP	Pos ↑	40.14	47.96	**+7.82**	70.34	76.62	**+6.28**	52.68	67.73	**+15.05**
		Neg ↓	33.34	36.63	+3.29	65.74	67.55	+1.81			
	SigLIP	Pos ↑	50.01	52.12	**+2.11**	80.20	80.81	**+0.61**	80.49	83.45	**+2.96**
		Neg ↓	40.00	38.51	**-1.49**	72.70	70.16	**-2.54**			
GradCAM	CLIP	Pos ↑	44.68	51.43	**+6.75**	74.94	79.80	**+4.86**	65.39	71.82	**+6.43**
		Neg ↓	33.83	34.23	+0.40	67.34	68.45	+1.11			
	SigLIP	Pos ↑	38.69	43.26	**+4.57**	71.43	73.61	**+2.18**	67.07	73.01	**+5.94**
		Neg ↓	34.78	37.21	+2.43	68.53	69.52	+0.99			
DAAM	SD 2	Pos ↑	65.71	66.34	**+0.63**	88.55	89.76	**+1.21**	83.07	85.48	**+2.41**
		Neg ↓	59.49	59.03	**-0.46**	86.39	87.33	+0.94			

evaluates the model's ability to correctly rank these attributions, it serves as a direct measure of our method's success in mitigating hallucination.

Our results indicate a significant improvement in AUROC, suggesting that OSP substantially enhances hallucination detection. Notably, this improvement does not come at the cost of localization performance, as the mIoU for positive prompts remains robust. Furthermore, the reduction in scores for the *negative prompts* demonstrates that inaccurate heatmaps are more effectively suppressed. Any improvement observed in the negative prompts, if present, is smaller than that of the positive prompts. In this set of experiments, increasing the gap between the positive and negative prompts was selected as the objective criterion. This trend is consistent across all dictionaries as shown in Table 2 (see Appendix D for further details). Importantly, OSP is not dependent on LLMs: dictionaries constructed from ImageNet class names or WordNet semantic relations already yield consistent improvements across all metrics. Furthermore, OSP is a computationally efficient operation, with a runtime of less than one second across all evaluated architectures and saliency methods. Detailed analysis of the computational overhead is provided in Appendix I.

5.3 Visual Results

Fig. 3, and figures in Appendix G present qualitative comparisons across all five attribution methods (AttentionCAM, CheferCAM, GradCAM, LeGrad, DAAM)

Table 2: Average performance per dictionary strategy on ImageNet-Segmentation. Values are averaged over all 9 method–model pairs (4 discriminative methods × 2 models + DAAM). Δ columns show the mean change introduced by OSP. For positive prompts (↑ desired), OSP should increase scores; for negative prompts (↓ desired), OSP should decrease them. Favorable changes are highlighted in **bold**.

Dictionary	Prompt	mIoU			mAP			AUROC		
		Base	+OSP	Δ	Base	+OSP	Δ	Base	+OSP	Δ
ImageNet Classes	Pos ↑	48.20	51.96	**+3.76**	77.79	80.32	**+2.53**	70.64	74.08	**+3.44**
	Neg ↓	42.09	41.02	**-1.07**	73.55	72.68	**-0.87**			
ImageNet + WordNet	Pos ↑	47.98	51.40	**+3.42**	77.79	80.32	**+2.53**	70.43	73.90	**+3.47**
	Neg ↓	40.12	41.20	+1.08	71.49	72.77	+1.28			
Gemini 3 Flash	Pos ↑	48.20	51.61	**+3.41**	77.79	80.41	**+2.62**	70.65	75.52	**+4.87**
	Neg ↓	40.12	40.27	+0.15	71.49	72.39	+0.90			
GPT-OSS 120B	Pos ↑	48.20	51.34	**+3.14**	77.79	80.29	**+2.50**	70.65	76.26	**+5.61**
	Neg ↓	40.12	39.90	**-0.22**	71.49	72.02	+0.53			

with and without OSP, spanning both discriminative (CLIP/SigLIP) and generative (Diffusion) architectures. Each figure shows heatmaps for the target objects present in the image as well as a non-existent distractor class, illustrating how standard methods produce hallucinated attributions for the absent concept while OSP effectively suppresses these ghost signals. While standard methods often hallucinate absent concepts, OSP suppresses these signals and improves **positive prompt** localization. Thus, OSP both mitigates hallucination and refines true attributions.

6 Practical Use Cases and User Study

6.1 Use Cases

A key implication of Theorem 1 (Linear Semantic Leakage) is that the severity of hallucination is predictive solely from the geometry of the text embedding space, without requiring access to image data or model inference. This mathematical property enables several practical applications for deploying Vision-Language Models in real-world interpretation tasks.

Fine-Grained Concept Disambiguation. A frequent failure in multimodal interpretability is the inability to distinguish between closely related semantic concepts or sub-categories. This is often linked to the grounding issues inherent in these models [18, 25]. This problem also affects Concept Bottleneck Models (CBMs) [12,18,27,37], where the concept extractor typically operates as a "black box." As shown in [6, 18] and illustrated in Fig. 4, standard attribution methods struggle to provide accurate heatmaps for these specific concepts. However, OSP rectifies these heatmaps, enabling robust, fine-grained concept disambiguation. We evaluate several attribution methods across various bird-related semantic concepts; additional accuracy results are provided in Appendix E. Ultimately, OSP can explain the concepts extracted by CBMs, making the previously opaque components of these models more transparent.

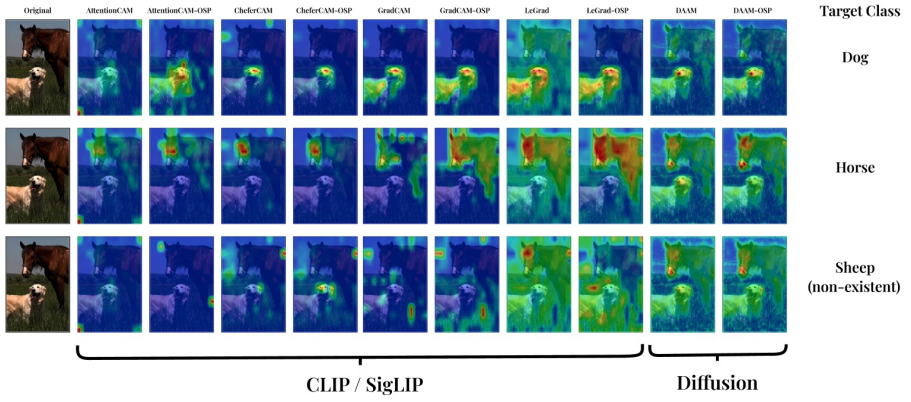

Fig. 3: Visual comparison across all methods on a **horse–dog** scene. Rows correspond to queries for "Dog" (target), "Horse" (target), and "Sheep" (non-existent distractor). OSP consistently eliminates hallucinated attributions for the absent sheep across both discriminative and generative methods.

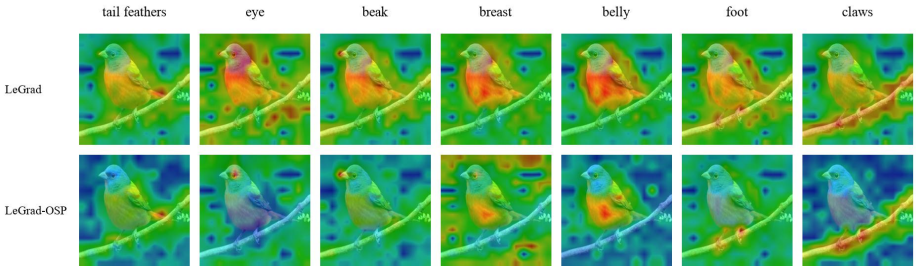

Fig. 4: Practical use case of OSP for fine-grained concept disambiguation. Standard methods suffer from semantic leakage and highlight the general object broadly, whereas OSP successfully disentangles the concepts to highlight specific features.

6.2 User Study

While our quantitative metrics validate OSP on ground-truth segmentation masks, a critical question remains: does OSP make attribution methods more *reliable* for human interpretation? To evaluate this, we conducted a user study with 200 participants, inspired by the human-aligned evaluation framework proposed by Kazmierczak et al. [19]. Participants were shown heatmaps generated by standard attribution methods, with and without OSP, and were asked: *"Based on the heatmap, which class is the model focusing on?"*. They then rated their confidence in their answer and, after the correct class was revealed, their trust in the explanation.

Results. The results are shown in Fig. 5 and Appendix F. OSP increases human identification accuracy, confidence, and trust scores across methods and

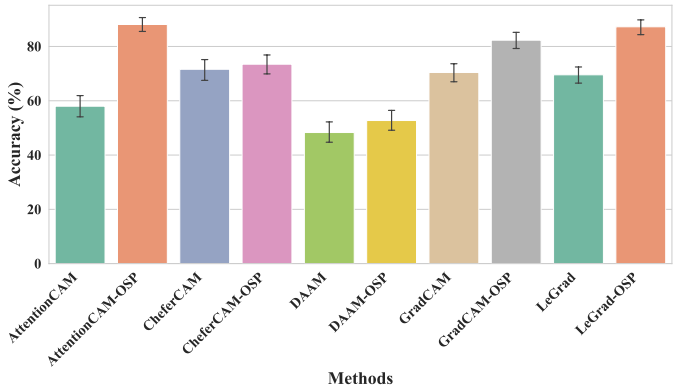

Fig. 5: User study results. Our method consistently achieves better accuracy, confidence, and trust scores.

architectures. This demonstrates that our geometric intervention not only provides a theoretically sound grounding but also makes attribution maps more reliable for end users, translating into measurable improvements in human interpretability. Detailed demographic distributions, statistical analysis including ANOVA results, deeper performance results including detailed confidence and trust scores, and subgroup analyses can be found in Appendix F. Extended results further confirm that OSP's improvements hold across subgroups.

7 Conclusion

We presented a geometric analysis of attribution hallucination in Vision-Language Models, showing that it arises from Linear Semantic Leakage: the unavoidable non-orthogonality of text embeddings in representation spaces. Building on this analysis, we introduced Orthogonal Semantic Projection, a modular geometric intervention that leverages Orthogonal Matching Pursuit to project query embeddings onto directions orthogonal to distractors. We proved that this orthogonalization mathematically eliminates ghost saliency signals by construction.

Extensive experiments across three foundational models, five attribution methods, four dictionary construction strategies, and two datasets demonstrate that OSP consistently improves AUROC while preserving or enhancing localization fidelity. A comprehensive user study further confirms that OSP yields gains in interpretability.

Looking ahead, we believe that the geometric perspective introduced here opens promising directions for extending OSP to larger-scale generative architectures and broader trustworthy AI applications. Beyond interpretability, the same orthogonalization principle could be directly applicable to high-precision image editing: by isolating a concept's unique semantic direction from shared distractors, OSP can localize and manipulate the target region while leaving correlated, non-target content untouched.

References

1. Black Forest Labs: FLUX.1. https://github.com/black-forest-labs/flux (2024), accessed: 2026-06-29
2. Bousselham, W., Boggust, A., Chaybouti, S., Strobelt, H., Kuehne, H.: LeGrad: An explainability method for vision transformers via feature formation sensitivity. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
3. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. arXiv preprint arXiv:2012.09838 (2020)
4. Chefer, H., Gur, S., Wolf, L.: Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 397–406 (2021)
5. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In: Proceedings of the International Conference on Learning Representations (ICLR) (2024)
6. Debole, N., Barbiero, P., Giannini, F., Passerini, A., Teso, S., Marconato, E.: If concept bottlenecks are the question, are foundation models the answer? arXiv preprint arXiv:2504.19774 (2025)
7. Elad, M.: Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing. Springer (2010)
8. Everingham, M., Eslami, S.M.A., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision **111**(1), 98–136 (2015)
9. Fel, T., Lubana, E.S., Prince, J.S., Kowal, M., Boutin, V., Papadimitriou, I., Wang, B., Wattenberg, M., Ba, D., Konkle, T.: Archetypal sae: Adaptive and stable dictionary learning for concept extraction in large vision models. arXiv preprint arXiv:2502.12892 (2025)
10. Gandelsman, Y., Efros, A.A., Steinhardt, J.: TextSpan: Interpreting CLIP's image representation via text-based decomposition. In: Proceedings of the International Conference on Learning Representations (ICLR) (2024)
11. Guillaumin, M., Kuttel, D., Ferrari, V.: Imagenet auto-annotation with segmentation propagation. International Journal of Computer Vision **110**(3), 328–348 (2014)
12. Havasi, M., Parbhoo, S., Doshi-Velez, F.: Addressing leakage in concept bottleneck models. Advances in Neural Information Processing Systems **35**, 23386–23397 (2022)
13. Hedström, A., Weber, L., Krakowczyk, D., Bareeva, D., Motzkus, F., Samek, W., Lapuschkin, S., Höhne, M.M.C.: Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. Journal of Machine Learning Research **24**(34), 1–11 (2023)
14. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems **43**(2), 1–55 (2025)
15. Huh, M., Cheung, B., Wang, T., Isola, P., Agrawal, P., Torralba, A.: The platonic representation hypothesis. arXiv preprint arXiv:2405.07987 (2024)
16. Kaul, P., Li, Z., Yang, H., Dukler, Y., Swaminathan, A., Taylor, C., Soatto, S.: Throne: An object-based hallucination benchmark for the free-form generations

of large vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27228–27238 (2024)
17. Kazerouni, A., Aghdam, E.K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., Merhof, D.: Diffusion models in medical imaging: A comprehensive survey. Medical Image Analysis **88**, 102846 (2023)
18. Kazmierczak, R., Azzolin, S., Berthier, E., Frehse, G., Franchi, G.: Enhancing concept localization in CLIP-based concept bottleneck models. Transactions on Machine Learning Research (2025)
19. Kazmierczak, R., Azzolin, S., Berthier, E., Hedström, A., Delhomme, P., Filliat, D., Franchi, G.: Benchmarking XAI explanations with human-aligned evaluations. arXiv preprint arXiv:2411.02470 (2024)
20. Li, X., Liu, Y., Dong, N., Qin, S., Hu, X.: PartImageNet++ dataset: Scaling up part-based models for robust recognition. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 396–414 (2024)
21. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 292–305 (2023)
22. Liang, W., Zhang, Y., Kwon, Y., Yeung, S., Zou, J.: Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 35, pp. 17612–17625 (2022)
23. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 740–755 (2014)
24. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
25. Liu, Y., Ji, T., Sun, C., Wu, Y., Zhou, A.: Investigating and mitigating object hallucinations in pretrained vision-language (clip) models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 18288–18301 (2024)
26. Nasiriany, S., Xia, F., Yu, W., Xiao, T., Liang, J., Dasgupta, I., Xie, A., Driess, D., Wahid, A., Xu, Z., Vuong, Q., Zhang, T., Lee, T.W.E., Lee, K.H., Xu, P., Kirmani, S., Zhu, Y., Ichter, B., Tompson, J., Hausman, K.: Pivot: Iterative visual prompting elicits actionable knowledge for VLMs. In: Proceedings of the International Conference on Machine Learning (ICML) (2024)
27. Oikarinen, T., Das, S., Nguyen, L.M., Weng, T.W.: Label-free concept bottleneck models. arXiv preprint arXiv:2304.06129 (2023)
28. Pach, M., Karthik, S., Bouniot, Q., Belongie, S., Akata, Z.: Sparse autoencoders learn monosemantic features in vision-language models. arXiv preprint arXiv:2504.02821 (2025)
29. Papadimitriou, I., Su, H., Fel, T., Kakade, S., Gil, S.: Interpreting the linear structure of vision-language model embedding spaces. arXiv preprint arXiv:2504.11695 (2025)
30. Park, K., Choe, Y.J., Veitch, V.: The linear representation hypothesis and the geometry of large language models. arXiv preprint arXiv:2311.03658 (2023)
31. Pati, Y.C., Rezaiifar, R., Krishnaprasad, P.S.: Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In: Proceedings of the Asilomar Conference on Signals, Systems and Computers. pp. 40–44 (1993)

32. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. In: Proceedings of the British Machine Vision Conference (BMVC) (2018)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 8748–8763 (2021)
34. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 4035–4045 (2018)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (2022)
36. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 618–626 (2017)
37. Srivastava, D., Yan, G., Weng, L.: Vlg-cbm: Training concept bottleneck models with vision-language guidance. Advances in Neural Information Processing Systems **37**, 79057–79094 (2024)
38. Tang, R., Liu, L., Pandey, A., Jiang, Z., Yang, G., Kumar, K., Stenetorp, P., Lin, J., Ture, F.: What the DAAM: Interpreting stable diffusion using cross attention. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). pp. 5644–5659 (2023)
39. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11975–11986 (2023)
40. Zhou, X., Liu, M., Yurtsever, E., Zagar, B.L., Zimmer, W., Cao, Y., Knoll, A.C.: Vision language models in autonomous driving: A survey and outlook. IEEE Transactions on Intelligent Vehicles **9**(1) (2024)