MIRAGE PROBES: HOW VISION MODELS FAKE VISUAL UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) can answer image-based questions confidently, and often correctly, even when no image is provided. This mirage behavior inflates benchmark scores without reflecting visual grounding. Prior work treats this as a single failure mode. We argue it is two. Using *Mirage Probes*, a contrastive probing framework that pairs paraphrased question variants with matched mirage and non-mirage labels on the same image, we show that mirage behavior is linearly decodable from image-present internal activations across four target sites in two open-source VLMs. A Naive Bayes text baseline fails to recover this signal, ruling out surface lexical confounds. Analysis of per-question text-only answerability and further latent space probing expose two distinct regimes: textual biases, where the model answers from language priors without engaging visual representations, and spurious images, where it constructs false visual content in latent space and answers as if grounded. The distinction has direct mitigation consequences: text-distribution cleaning can address the first regime but cannot reach the second, since spurious-image mirages live in the model's visual representations rather than its text. Faithful visual grounding will require interventions at the representational level. Code is available here.

1 INTRODUCTION

Vision-language models (VLMs) are increasingly used for tasks that require faithful visual understanding, including visual question answering (Antol et al., 2015; Kafle & Kanan, 2017), document analysis (Mathew et al., 2021; Kim et al., 2022), scientific reasoning (Lu et al., 2022), and medical image interpretation (Li et al., 2023a; Moor et al., 2023; Zhang et al., 2023). In such settings, a correct answer is insufficient unless it is grounded in the visual input. Yet recent work on mirage behavior (Asadi et al., 2026) challenges the assumption that benchmark success reflects faithful visual reasoning: VLMs can answer image-based questions confidently, and sometimes correctly, even when the image is absent. This suggests that performance on multimodal benchmarks such as MMMU (Yue et al., 2024) and VQA-RAD (Lau et al., 2018) may partly reflect textual priors, dataset regularities, or benchmark-specific shortcuts rather than genuine use of visual evidence.

To remediate this behavior, we must understand what gives rise to it internally. Prior work largely treats mirages as a single failure mode (Asadi et al., 2026), but the same output-level behavior may be facilitated by different latent-space mechanisms. A model may form internal representations consistent with unsupported visual content, effectively behaving as though a *spurious image* were present. Alternatively, it may answer directly from *textual biases* (benchmark-specific or distributional priors) without forming any false visual representation. These mechanisms, illustrated in Figure 1, imply different interventions: text-distribution cleaning may reduce shortcut-driven mirages, but may be insufficient when mirage behavior is associated with unsupported visual inference.

In this work, we introduce *Mirage Probes*, a representation-level framework for studying mirage behavior in VLMs. We construct datasets of mirage-associated and non-mirage-associated image-conditioned responses by comparing model generations with and without visual input, and we build contrastive pairs from lightly paraphrased question variants that preserve image-question semantics while yielding different mirage labels, reducing surface-level textual confounds. We then probe internal activations from open-source VLMs at four sites (namely, residual stream states, MLP outputs, post-attention outputs, and individual attention-head outputs), using probe families that test whether

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

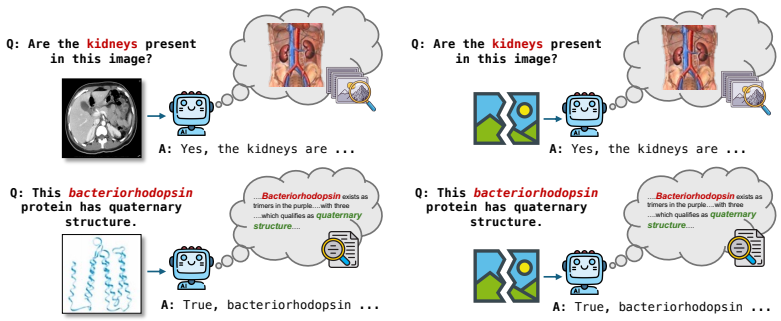


Figure 1: **Two distinct mirage mechanisms.** VLMs seem to exhibit two different kinds of mirage behavior, *spurious images* and *textual biases*, depicted here. Each row shows a mirage response to the same question with and without an image present. In the first example, the model achieves mirage behavior by building and referring to a false visual representation. Differently, in the second case, the model does not utilize visual information at all, relying solely on textual priors provided by the rich question distribution to reach the correct response.

mirage information is encoded as a single linear direction, requires a nonlinear transformation, or is most cleanly recovered from relational structures.

Our results show that mirage behavior is linearly decodable from VLM latent representations across all four activation sites when images are provided, indicating that it is not merely an image-absent or output-level artifact. Two-layer MLP probes do not meaningfully outperform linear ones, and contrastive difference probes recover the signal most cleanly. Together, these results are consistent with mirage information being encoded along recoverable linear directions rather than requiring nonlinear extraction. A Naive Bayes textual baseline trained on response text is consistently weaker than our contrastive probes, suggesting that this signal is not reducible to surface lexical features.

Examining image-absent model responses, we notice that when rich textual priors are available, mirage-like answers take the form of confident and correct explanations that do not reference any visual information at all. To test whether or not these lexical mirages are facilitated by a distinct mechanism, we train latent space probes to separate these examples from those involving fabricated visual content. We find that linear probes achieve high separability in VLM latent space, compared to weaker accuracy on semantic and lexical baselines. Accordingly, we propose, as a hypothesis motivating further mechanistic work, that behavioral mirages conflate at least two representational regimes – spurious images and textual biases – rather than constituting a single phenomenon. More broadly, our findings raise concerns for VLM training and evaluation: reducing textual shortcuts is necessary, but may not be sufficient to ensure faithful visual grounding.

Our contributions are:

1. **We show mirage behavior is linearly decodable from image-present latent space.** Linear probes distinguish mirage-associated from non-mirage-associated image-conditioned generations across multiple activation sites, and contrastive difference probes recover this signal most cleanly.

2. **We introduce *Mirage Probes*, a contrastive probing framework for mirage behavior.** By pairing semantically similar question variants with different mirage labels, our framework reduces surface-level textual confounds, as confirmed by comparison against a Naive Bayes textual baseline.

3. **We hypothesize two distinct mirage mechanisms.** Mirage regime probe accuracies suggest that prior behavioral definitions group together *spurious-image* and *textual-bias* mirages, though we leave causal validation of this distinction to future work.

4. **We argue that text-only mitigation may be insufficient.** Text-distribution cleaning targets textual-bias mirages but is unlikely to address spurious-image mirages, since the latter are associated with internal representations rather than surface textual patterns.

2 Related work

Hallucination in vision-language models. Hallucination in VLMs has been studied extensively, primarily as the assertion of visual content not present in the input image. Object hallucination benchmarks such as POPE (Li et al., 2023b) and CHAIR (Rohrbach et al., 2018) probe whether models invent objects, while subsequent work has extended the analysis to attribute and relational hallucinations (Sun et al., 2024; Jiang et al., 2024). Recent works provide broad treatments of the phenomenon, its causes, and mitigations (Liu et al., 2024; Bai et al., 2024; Goldfeder et al., 2026).

Image-absent generation and modality bypass. A growing line of work examines cases in which VLMs appear to bypass visual input. Tong et al. (2024) and Li et al. (2024) show that state-of-the-art VLMs frequently fail on visual questions that humans answer trivially. Lee et al. (2025) find that VLMs answer many questions incorrectly due to an over-reliance on language priors. Hua et al. (2025) study VLM behavior when modalities present conflicting information. Furthermore, Yin et al. (2026) demonstrate that truthful inference and visual perception rely on different subsets of attention heads, and Rudman et al. (2026) show that ablating the correct subset of heads can reduce prompt-induced hallucination. Most directly, Asadi et al. (2026) introduce the term *mirage* for the observation that VLMs produce confident, often correct answers to image-grounded questions even when no image is provided, and hypothesize that the same mechanism may operate when images are present. We aim to test this hypothesis at the level of internal representations.

Probing and mechanistic interpretability. Linear probes on intermediate representations have a long history in NLP interpretability, from early diagnostic classifiers (Alain & Bengio, 2016; Belinkov, 2022) to syntactic probes (Hewitt & Manning, 2019) and the broader study of what linear directions encode in language model representations (Tenney et al., 2019; Park et al., 2023; Lomasov et al., 2025; Theodoridis et al., 2026). Contrastive probe constructions (Burns et al., 2024; Marks & Tegmark, 2024) have proven particularly effective at recovering features otherwise masked by surface confounds, and we adopt a contrastive-pair design for the same reason. We extend this line of work by targeting a specific behavioral failure, mirage generation, and showing that it has a recoverable latent representation that further admits a two-mechanism decomposition.

3 Methods

3.1 Mirage Behavior

Let M denote a vision-language model. For each example i, let x_i be the image, q_i be the textual question, and let

$$r_i^{\text{img}} = M(x_i, q_i), \qquad r_i^{\emptyset} = M(\emptyset, q_i)$$

denote the model response with and without the image, respectively. Mirage behavior occurs when the model responds confidently, and sometimes correctly, even when the image is absent. In this case, the model may appear to use visual evidence, while actually relying on textual priors, dataset regularities, or unsupported internal visual assumptions.

We define a binary mirage label $y_i \in \{0, 1\}$ for each with-image response. The label $y_i = 1$ denotes a mirage-like response, and $y_i = 0$ denotes a non-mirage-like response. Since a with-image mirage cannot be directly observed from the output alone, this label is treated as a behavioral proxy obtained by comparing r_i^{img} and $r_i^{\emptyset}$. The goal of our probing setup is to test whether this behavioral label is recoverable from the model's internal representations.

3.2 Dataset Construction

Annotation Scheme. Our goal is to assemble labeled examples of mirage and non-mirage with-image responses. Whether a given with-image generation is actually a mirage cannot be determined from the output alone, so we rely on a heuristic. Following the hypothesis from Asadi et al. (2026) that distribution shift between with-image and without-image responses signals genuine reliance on provided visual input, we adopt the following labeling scheme.

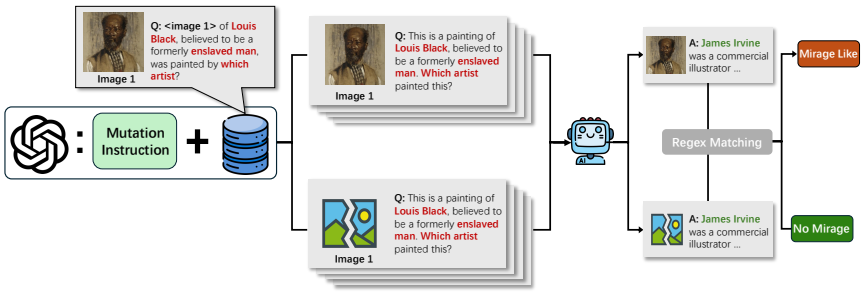


Figure 2: **Contrastive dataset construction overview.** In order to produce contrastive pairs, we mutate each base question four times with GPT-4o-mini and generate with and without-image responses to each variant. Then, for each pair of responses, we use cosine similarity and a regex annotator to identify mirage and non-mirage behavior. If a base question's group contains at least one mirage and non-mirage pair, we store each pair's with-image response to form a contrastive example.

A response is **mirage-like** if the with-image and without-image generations produce the same answer and have a bag-of-words vector cosine similarity above 0.7. A response is **non-mirage-like** if the without-image generation explicitly expresses uncertainty due to missing visual context or states that answering is impossible without the image, both detected through regex checks. A response is **ambiguous** if the without-image generation is itself mirage-like but the with-image answer differs in either answer or content similarity. We choose to include an explicit ambiguous label because in such cases, it is impossible to determine whether the with-image generation represents a legitimate use of visual information or is just a different mirage answer triggered by the distribution shift of including the image.

We also tried two alternative schemes. Folding ambiguous responses into the mirage-like class and labeling non-mirage examples with the B-Clean criterion described by Asadi et al. (2026) (questions that cannot be answered correctly without the image are considered safe), both reduced probe accuracy substantially. Thus, we select the scheme above for our main experiments.

Benchmarks. We draw questions from four widely-used visual question answering benchmarks chosen to span medical and general expertise, along with free-form and multiple-choice formats. **VQA-RAD** (Lau et al., 2018) contains clinically generated free-form and yes/no questions about radiology images. **MMMU-Pro** (Yue et al., 2025) is a multiple-choice expert-level reasoning benchmark spanning college-level subjects across many disciplines, designed to be more robust to text-only shortcut solutions than its predecessor (Yue et al., 2024). **MedXpertQA** (Zuo et al., 2025) provides expert-level multiple-choice medical questions targeted at reasoning rather than recall. **MicroVQA** (Burgess et al., 2025) consists of multiple-choice questions over biomedical microscopy images.

Contrastive Pairs. Linear probes benefit substantially from training on contrastive pairs, where positive and negative examples differ only in the property of interest (Burns et al., 2024; Marks & Tegmark, 2024). Without contrastive structure, surface-level features (topic words, response length, question style) often correlate with labels, causing probes to learn unintended shortcuts rather than the target representation.

For each base question, we generate four mutations using GPT-4o-mini (Hurst et al., 2024). A mutation rephrases or rearranges the question while preserving its meaning and adding no new information. The intent is mild distributional perturbation, sufficient to flip behavior between mirage and non-mirage for some questions while keeping the image and the underlying intent fixed. When the responses to a base question and its four mutations together contain at least one mirage-like and one non-mirage-like response, we draw a single contrastive pair from that question set. We never draw more than one pair per base question to preserve diversity.

We construct two datasets per evaluated model. The **contrastive-pairs** dataset contains only such pairs, with the same image and near-identical question content across both classes. The **all-examples**

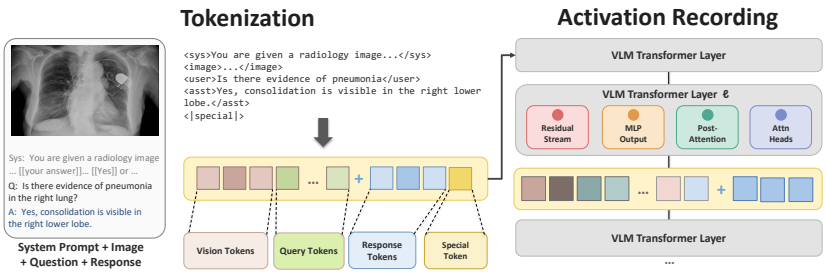


Figure 3: **Activation extraction overview.** The VLM first generates a response to a user query, following system prompt instructions and referring to attached images if any are provided. The full conversation, model response included, is then chat-template tokenized and fed into the VLM for extraction. Any images are first processed by the model's vision encoder and then projected into vision tokens, which are prepended to the user's query. Then, the model conducts an LLM decoder-only pass over the full conversation, and we extract activations from the residual stream, MLP outputs, post-attention outputs, and individual attention head outputs at each layer.

dataset contains every response annotated as mirage-like or non-mirage-like, capped at one example per class per base question. The contrastive split aims to test whether the mirage signal survives without surface-level confounds, while the all-examples split contains a larger quantity of examples, which may aid generalization if textual confounds do not override the desired direction. Our dataset construction procedure is summarized in Figure 2.

3.3 Model Architecture and Activation Extraction

We probe three open-source vision-language models spanning different model sizes and architectures. **Ovis2.5-2B** (Lu et al., 2024) is a 2B-parameter VLM that aligns vision and language representations through a structured embedding scheme. **GLM-4.6-flash** (Team GLM et al., 2024) is the smaller variant in the GLM multimodal series. **Qwen3-32B-VL-Instruct** (Bai et al., 2025) is the 32B-scale vision-language model in the Qwen3 family. Standard VLM design glues a vision encoder to a language model through a learned projection that maps image patch features into the language model's embedding space (Li et al., 2022; Liu et al., 2023). In our experiments, image tokens are prepended to the user's textual query, after which the language model performs standard causal attention.

We extract activations at four targets within each model: the **residual stream** at every layer, the **MLP output** at every layer, the **post-attention output** at every layer, and **every individual attention head output** at every layer. We initially included the vision encoder and the projection layers but found their representations carried no detectable mirage signal under any probing strategy and accordingly dropped them from our main analysis.

For each model and each annotated dataset, we run the model on chat-formatted conversations and cache activations under three pooled token-position aggregations, building on the position scheme introduced by Theodoridis et al. (2026). The first aggregation, *text-nonspecial-mean*, averages activations over all non-special text tokens. The second, *vision-tail*, averages over all vision tokens and concatenates their pooled vector with the last special token in the response. The third, *vision-text-mean*, averages over all vision tokens and all non-special text tokens jointly. Our activation extraction procedure is illustrated in Figure 3.

Final per-model dataset sizes are presented in Appendix A. GLM-4.6-flash produces too few non-mirage responses on three of four benchmarks to support stable probing, so we drop it from the main results. MicroVQA yields too few non-mirage responses for stable probing across models, so it is also dropped from the main results. Qualitatively, we note that MicroVQA has the lowest visual reliance among the four benchmarks, which is likely related to its extremely high associated mirage rate. Ovis2.5-2B produces a relatively small viable dataset for some benchmarks, but we retain the model nonetheless, as consistent results across Ovis2.5-2B and Qwen3-32B-VL-Instruct rule out scale-specific artifacts and increase the robustness of our findings.

Table 1: **Residual and attention probing results.** We train four probing strategies on our contrastive and all-examples datasets, filtered to each individual benchmark first and then all at once. Results targeting the residual stream are presented on the left, while results on attention heads and post-attention outputs are shown on the right. Our results indicate that mirage behavior is linearly represented in image-present latent space, with consistent non-trivial accuracy across benchmarks and targets.

Residual

Strategy	VQA-RAD	MMMU-Pro	MedXpertQA	All
Ovis – Contrastive				
LogReg	**75.00%**	65.00%	68.00%	72.20%
MLP	72.20%	62.60%	64.00%	**72.80%**
Concat	**67.00%**	58.20%	52.00%	64.00%
Diff	**97.40%**	90.80%	71.00%	91.80%
Ovis – All Examples				
LogReg	68.09%	**73.10%**	57.14%	71.36%
MLP	67.23%	**75.17%**	60.00%	70.12%
Concat	58.72%	64.14%	51.43%	**68.64%**
Diff	**98.30%**	88.97%	57.14%	92.10%
Qwen – Contrastive				
LogReg	**72.00%**	57.80%	57.20%	61.20%
MLP	**72.40%**	59.60%	57.40%	60.40%
Concat	**67.40%**	53.40%	50.60%	57.00%
Diff	**96.00%**	79.40%	76.20%	77.40%
Qwen – All Examples				
LogReg	61.43%	**65.06%**	63.17%	59.04%
MLP	73.80%	**76.35%**	70.43%	72.69%
Concat	70.14%	**72.87%**	68.70%	70.41%
Diff	**95.77%**	80.70%	76.52%	84.38%

Attention

Strategy	VQA-RAD	MMMU-Pro	MedXpertQA	All
Ovis – Contrastive				
LogReg	**74.67%**	68.00%	70.00%	73.67%
MLP	**77.33%**	69.00%	73.33%	71.00%
Concat	**75.67%**	66.33%	66.67%	72.33%
Diff	87.67%	**91.33%**	71.67%	85.00%
Ovis – All Examples				
LogReg	75.18%	77.01%	**85.71%**	74.07%
MLP	75.18%	75.86%	**90.48%**	73.25%
Concat	73.76%	71.26%	**80.95%**	73.25%
Diff	**90.78%**	87.36%	90.48%	88.89%
Qwen – Contrastive				
LogReg	**81.00%**	63.67%	69.33%	64.33%
MLP	**75.67%**	66.67%	71.67%	64.00%
Concat	**75.33%**	64.33%	65.00%	61.33%
Diff	**95.67%**	81.67%	83.33%	82.33%
Qwen – All Examples				
LogReg	69.05%	**69.62%**	67.07%	61.52%
MLP	**75.12%**	75.07%	74.88%	73.97%
Concat	73.71%	**75.94%**	71.01%	73.52%
Diff	**95.77%**	82.32%	78.74%	85.39%

3.4 Latent Space Probes

We train four families of probes, each motivated by a different prior on how mirage information could be encoded. Let $h_{\ell,i} \in \mathbb{R}^{d_\ell}$ denote the activation of example i extracted from layer or component ℓ. In our setting, ℓ may correspond to a residual stream state, an MLP output, a post-attention output, or an individual attention-head output. A probe is a classifier g_ϕ trained to predict the mirage label from the activation:

$$\hat{y}_i = g_\phi(h_{\ell,i}).$$

Thus, probe accuracy measures whether mirage behavior is decodable from the model's latent space.

Logistic regression is a standard linear probing strategy (Alain & Bengio, 2016; Goldowsky-Dill et al., 2025). It predicts

$$p(y_i = 1 \mid h_{\ell,i}) = \sigma(w^\top h_{\ell,i} + b),$$

where w and b are learned parameters and σ is the sigmoid function. Above-chance accuracy implies the target property is linearly decodable from the model's representation space, which is the cleanest evidence that the model encodes that property as a feature direction.

Two-layer MLP probes test whether mirage information is decodable through a lightweight nonlinear transform but not as a single linear separator. A gap between linear and MLP accuracy points to distributed or nonlinear encoding.

Logistic regression on concatenated activations trains a linear probe on the concatenation of activations from all model layers at once. If the model has L layers, we define

$$\bar{h}_i = [h_{1,i}; h_{2,i}; \ldots; h_{L,i}],$$

and train logistic regression on $\bar{h}_i$. Such probes allow us to determine whether mirage behavior is sparsely represented across layers instead of cleanly identifiable at a specific point.

Logistic regression on the difference of activations trains on the elementwise difference between with-image and without-image activations for each example. Let $h_{\ell,i}^{\text{img}}$ denote the activation when the image is provided, and let $h_{\ell,i}^{\emptyset}$ denote the activation when the image is removed. We define

$$\Delta h_{\ell,i} = h_{\ell,i}^{\text{img}} - h_{\ell,i}^{\emptyset}.$$

A linear probe trained on $\Delta h_{\ell,i}$ tests whether the representational shift caused by adding the image separates mirage-like responses from non-mirage-like responses.

Table 2: **Naive Bayes classifier.** In order to determine whether or not our probes are simply picking up on surface level confounds, we train a text-based Naive Bayes classifier. The model's performance indicates that our contrastive probes pick up on deep mirage mechanisms, while all-examples probes are likely distracted by surface-level features.

Model (Setting)	VQA-RAD	MMMU-Pro	MedXpertQA	All
Ovis (Contrastive)	61.88%	57.67%	**66.67%**	64.85%
Ovis (All Examples)	75.38%	**81.11%**	67.50%	75.74%
Qwen (Contrastive)	**59.08%**	54.80%	45.45%	52.24%
Qwen (All Examples)	77.08%	**81.50%**	79.64%	72.80%

3.5 Experimental Setup

We train all four types of latent space probes on per-layer activations generated with Ovis and Qwen on conversations from our contrastive and all-examples datasets. Data points where the with-image response is fewer than 10 tokens long are excluded. For each combination of probing strategy, model, and dataset, we conduct four training runs, one on examples exclusively from each of our three retained benchmarks and one on all three benchmarks combined. MLP probes are trained with 2 layers, a hidden dimension of 512, and a GELU activation function. All probes are trained with an Adam optimizer and a learning rate of 0.03, and features are train-set normalized before training.

Residual stream probes. When targeting the residual stream, we aggregate activations with all three options (*text-nonspecial-mean*, *vision-tail*, *vision-text-mean*). Contrastive probes are trained using a 90/10 train/valid split (approximated with hard-coded values for Ovis) and then evaluated on a held-out class-balanced, benchmark-stratified dataset that consist of up to 100 data points from the all-examples dataset. This procedure allows us to maintain robust evaluation while also using our limited contrastive data to the greatest effect. All-examples probes are trained using a 70/10/20 train/valid/test split. We train probes on 5 seeded splits with 3 random initializations per seed and a 4-option L2 regularization hyperparameter sweep (0, 1, 10, 100) per random initialization, resulting in 12 probes trained per seed. The best of the 12 probes, selected through early stopping on validation accuracy (max 800 epochs), is tested, either once on the held-out portion of the split in the all examples setting or averaged over an additional 5 held-out dataset seeds in the contrastive setting. This procedure is repeated independently for each aggregation strategy. We take our final accuracy for one experimental configuration as the averaged held-out test accuracy across the 5 seeded splits for the aggregation strategy that performed best.

MLP and Attention Probes. In order to keep our experiments tractable while adding a huge number of additional targets, we slightly modify the above procedure. First, we only aggregate activations with *text-nonspecial-mean*, as this strategy generally performs best in the residual setting. Next, we remove the 3 random initializations per seed, drop the number of split seeds to 3, and drop the number of held-out contrastive setting test seeds to 3. All other aspects of the experimental setup are the same. For attention probes, we report the best accuracy for each experimental configuration as the best held-out test accuracy across all of the attention head and post-attention outputs at all layers.

4 Results

4.1 Mirage Mechanisms Linearly Represented in Image-Present Latent Space

As presented in Table 1, we demonstrate that mirage behavior is clearly represented across model layer components through non-trivial probing accuracy. When targeting the residual stream, in Ovis's contrastive and all-examples settings, we achieve linear probe accuracies of up to 75% and 73%, respectively. Similarly, in Qwen's contrastive and all-examples settings, we achieve linear probe accuracies of up to 72% and 65%. Attention head and post-attention targets show similar results, with Ovis accuracies of up to 74% and 85% and Qwen accuracies of up to 81% and 69%, respectively. Generally, across all contrastive settings, multi-layer perceptron and concatenated-activations probes do not significantly outperform logistic regression, indicating that mirage mechanisms are represented linearly in latent space. We note that difference-in-activations probes achieve remarkably high accuracy, motivating their use as mirage-behavior classifiers. Probe strategy results

Table 3: **Image reliance per benchmark.** We use GPT-5-mini to annotate perceived image reliance across three benchmarks: VQA-RAD, MMMU-Pro, and MedXpertQA. We find that VQA-RAD is the only dataset wherein questions that generate mirage responses are consistently classified as image-reliant, supporting a notion of a distinct, spurious-image-based mirage mechanism.

VQA-RAD			**MMMU-Pro**			**MedXpertQA**		
	+ Mirage	- Mirage		+ Mirage	- Mirage		+ Mirage	- Mirage
Image-Free	60	3	Image-Free	836	18	Image-Free	2399	15
Image-Reliant	2909	236	Image-Reliant	1139	1421	Image-Reliant	2413	260

on MLP targets are presented in Appendix E. We confirm that this finding generalizes across five models and five benchmarks in Appendix G and provide causal verification in Appendix H.

4.2 Mirage Probes Framework Effectively Reduces Surface Level Confounds

We train a Naive Bayes textual classifier following a similar procedure to that used for probe training in order to determine whether or not our probes are detecting surface-level patterns or deeper mirage mechanisms. Detailed training details are discussed in Appendix B, and results are presented in Table 2. Overall, we find that our contrastive *Mirage Probes* framework significantly reduces surface-level separability due to its low semantic and topical variance across class examples. In particular, we note that contrastive setting accuracy ranges from 45% to 66% while all-examples setting accuracy ranges from 67% to 81%, emphasizing the reduction in confounds our framework is able to achieve. Crucially, contrastive logistic regression probe accuracy remains higher than Naive Bayes accuracy across all models, benchmarks, and targets, indicating with high confidence that our probes pick up on deeper mechanistic features. However, we note that increases in probe accuracy on the MMMU-Pro and MedXpertQA benchmarks in the all-examples setting are directly and consistently correlated with sharp increases in surface-level separability, indicating that these probes are likely picking up on textual cues instead of identifying the desired mirage mechanism. Further analysis of per-dataset textual confounds is presented in Appendix D.

4.3 Two Distinct Mirage Mechanisms: Spurious Images and Textual Biases

While developing our method, we noticed that many questions, particularly in the multiple choice benchmarks, could potentially be answered purely from the provided text, without needing additional insight from an image. We quantify this trend in Table 3. Specifically, we use a few-shot prompted version of GPT-5-mini (Singh et al., 2025) to annotate each benchmark's questions for perceived image dependence, marking questions that a human could not answer without referencing the provided visual material as image-reliant, and questions that a human could answer from the text alone as image-free. We find that across both MMMU-Pro and MedXpertQA, many image-free examples exist, and mirage-like generations are relatively evenly split between image-reliant and image-free questions, leading us to wonder whether or not mirages are enabled by the same underlying mechanisms in these distinct settings.

Image-reliant questions tend to produce mirage-like responses that clearly fabricate false visual information, while image-free mirages tend to answer confidently without referencing visual information at all. In order to determine whether this behavioral difference is supported by two separate latent mechanisms or rather a context-dependent instantiation of the same mechanism, we again employ linear probes. We use Claude-Sonnet-5 (Anthropic, 2026) agents to label Ovis and Qwen mirage-like image-absent generations as image-fabricating or lexical, select clean, label-balanced datasets of 52 to 348 examples per configuration, create 70/30 train/test splits, and train L2 regularized ($\lambda = 1$) logistic regression probes on VLM residual activations, all-mpnet-base-v2 (Song et al., 2020) residual activations, and TF-IDF vectors. Mean results over 5 seeds are presented in Table 4.

Across all six configurations, we find that probes trained on Ovis and Qwen activations significantly outperform controls, with increases of up to 12.8% and 12.1%, respectively, indicating that VLM-specific mechanisms are responsible for the high separability and thus suggesting that two distinct underlying regimes enable the two types of observed behavioral mirages. Given the contexts in

Table 4: **Two mechanism probe accuracies.** We train logistic regression probes to separate visual-fabrication mirages from lexical mirages using VLM, mpnet, and TF-IDF embeddings. We train probes on Ovis and Qwen responses from MMMU-Pro, MedXpertQA, and both benchmarks combined. Across all six configurations, probes trained on VLM embeddings outperform controls.

	Ovis Probe Accuracy			Qwen Probe Accuracy		
Benchmark	VLM	mpnet	TF-IDF	VLM	mpnet	TF-IDF
MMMU-Pro	**70.4**	66.7	64.1	**78.8**	71.2	65.0
MedXpertQA	**73.6**	60.8	60.0	**78.2**	73.6	69.1
Combined	**68.3**	62.7	58.8	**79.5**	67.4	56.3

which they arise, we name the mechanisms *spurious images* and *textual biases*. We define the usage of spurious images as the use of information pulled from visual priors, or in other words cases where the model constructs false images in latent space in order to answer the user's query. Differently, we define the usage of textual biases as the use of purely textual priors to answer the provided question, with the model ignoring the presence of visual information altogether.

We hypothesize that these two distinct mechanisms are natural consequences of VLM reward hacking (Skalse et al., 2025) during the reinforcement learning process and should thus respectively occur when they represent the path of least resistance to answering a question correctly. More concretely, we believe that textual priors will be used to answer any question wherein the model is able to determine a single answer with high confidence without any access to visual information. Spurious images will only be created when the provided textual distribution is not sufficiently rich to answer the question but is instead able to evoke false visual priors that can lead to a correct answer. Importantly, the directions found by our regime probes are post-hoc readout directions (null causal validation in Appendix J), so support for our two mechanism hypothesis remains correlational. We hope these findings provide motivation for further mitigation-oriented interpretability work.

5 Conclusion

We introduce *Mirage Probes*, a representation-level framework for studying mirage behavior in VLMs. Across two open-source VLMs and three VQA benchmarks, we find that mirage-associated generations are linearly decodable from image-present internal activations across residual stream, MLP, post-attention, and attention-head sites. Contrastive difference probes recover this signal most cleanly, while two-layer MLP probes do not meaningfully outperform linear probes, suggesting that mirage information is encoded along recoverable linear directions rather than requiring nonlinear extraction. A Naive Bayes text baseline remains weaker, indicating that the contrastive probe signal is not reducible to surface lexical features.

Our results further suggest that behavioral mirages are not a single phenomenon. Text-only answerability analysis and further probing results are consistent with two regimes: *spurious images*, where the model behaves as though unsupported visual content were present in latent space, and *textual biases*, where it answers from text-distributional priors without engaging visual representations. We view this decomposition as an empirically motivated hypothesis, not a causal claim, and validating it will require more fine-grained interventions. The mirage mechanism distinction matters for mitigation: text-distribution cleaning may reduce textual-bias mirages, but spurious-image mirages likely require representational interventions. The *Mirage Probes* framework provides a starting point for diagnosing and ultimately improving faithful visual grounding in VLMs.

6 Limitations

Several limitations qualify our findings. First, our mirage and non-mirage labels rely on heuristic criteria, which introduce some minimal label noise. Second, our two-mechanism account (spurious images versus textual priors) is based on correlation evidence and should be read as a hypothesis motivating further mechanistic work rather than a settled finding.

ETHICS STATEMENT

As diagnostic interpretability research, this work does not introduce new generative capabilities. Its primary positive impact is enabling more faithful evaluation of VLMs in safety-critical domains such as medical image interpretation, where ungrounded mirage responses pose direct risks to users. We do not see a direct path to negative societal impact beyond risks already present in the underlying open-source models we probe.

REPRODUCIBILITY STATEMENT

We release the codebase and data associated with this work with the submission (link in Abstract), along with a detailed explanation of how to reproduce the main results, including environment setup, dataset loading, and probing. We also describe our experimental setup in great detail in Section 3.5.

AI USAGE STATEMENT

LLM coding agents were used to generate scripts for this work. We provided detailed specifications for all scripts before generating. All implementation details and results were sanity checked manually by humans. LLMs were used minimally to draft some small portions of the final paper, which were then revised by humans.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Anthropic. System card: Claude sonnet 5, 2026. URL https://www-cdn.anthropic.com/9e6a1044980d8c4ed85669faf9c2a8342e2e9f1e/Claude%20Sonnet%205%20System%20Card.pdf. Accessed: 2026-09-25.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

Mohammad Asadi, Jack W O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Rajabalifardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. Mirage: The illusion of visual understanding. *arXiv preprint arXiv:2603.21687*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

James Burgess, Jeffrey J Nirschl, Laura Bravo-Sánchez, Alejandro Lozano, Sanket Rajan Gupte, Jesus G Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha Bhowmik, Zachary Coman, et al. Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19552–19564, 2025.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision, 2024. URL https://arxiv.org/abs/2212.03827.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Preetam Prabhu Srikar Dammu, Hayoung Jung, Anjali Singh, Monojit Choudhury, and Tanu Mitra. "they are uncultured": Unveiling covert harms and social threats in llm generated conversations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20339–20369, 2024.

Judah Goldfeder, Philippe Wyder, Yann LeCun, and Ravid Shwartz Ziv. Ai must embrace specialization via superhuman adaptable intelligence. *arXiv preprint arXiv:2602.23643*, 2026.

Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. Detecting strategic deception using linear probes. *arXiv preprint arXiv:2502.03407*, 2025.

Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of llms. *Advances in neural information processing systems*, 37:8093–8131, 2024.

Michael Hanna, Sandro Pezzelle, and Yonatan Belinkov. Have faith in faithfulness: Going beyond circuit overlap when finding model mechanisms. *arXiv preprint arXiv:2403.17806*, 2024.

John Hewitt and Christopher D Manning. 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)*, pp. 4129–4138, 2019.

Tianze Hua, Tian Yun, and Ellie Pavlick. How do vision-language models process conflicting information across modalities? *arXiv preprint arXiv:2507.01790*, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27036–27046, 2024.

Kushal Kafle and Christopher Kanan. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 163:3–20, 2017.

Navreet Kaur, Hoda Ayad, Hayoung Jung, Shravika Mittal, Munmun De Choudhury, and Tanu Mitra. Who's asking? simulating role-based questions for conversational ai evaluation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 35508–35541, 2026.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pp. 498–517. Springer, 2022.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251, 2018.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. Vlind-bench: Measuring language priors in large vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4129–4144, 2025.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. *Advances in Neural Information Processing Systems*, 37:17044–17068, 2024.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564, 2023a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023b.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Semyon Lomasov, Judah Goldfeder, Mehmet Hamza Erol, Matthew So, Yao Yan, Addison Howard, Nathan Kutz, and Ravid Shwartz Ziv. Exploring human-ai conceptual alignment through the prism of chess. *arXiv preprint arXiv:2510.26025*, 2025.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521, 2022.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, 2024.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2024. URL https://arxiv.org/abs/2310.06824.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine learning for health (ML4H)*, pp. 353–367. PMLR, 2023.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, 2018.

William Rudman, Michal Golovanevsky, Dana Arad, Yonatan Belinkov, Ritambhara Singh, Carsten Eickhoff, and Kyle Mahowald. Mechanisms of prompt-induced hallucination in vision-language models. *arXiv preprint arXiv:2601.05201*, 2026.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking, 2025. URL https://arxiv.org/abs/2209.13085.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33: 16857–16867, 2020.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, 2024.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Jingyu Sun, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024. URL https://arxiv.org/abs/2406.12793.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4593–4601, 2019.

Nikos Theodoridis, Reenu Mohandas, Ganesh Sistu, Anthony Scanlan, Ciarán Eising, and Tim Brophy. Probing visual concepts in lightweight vision-language models for automated driving, 2026. URL https://arxiv.org/abs/2603.06054.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9568–9578, 2024.

Jianghao Yin, Qin Chen, Kedi Chen, Jie Zhou, Xingjiao Wu, and Liang He. Dynamic multimodal activation steering for hallucination mitigation in large vision-language models. *arXiv preprint arXiv:2602.21704*, 2026.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. pp. 15134–15186, 2025.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025.

APPENDIX

A DATASET SIZES

Table 5: Post-filtering dataset sizes for contrastive and all-examples probe datasets.

Model	Benchmark	Contrastive Samples	All-Examples Samples
Ovis	VQA-RAD	98	235
Ovis	MMMU-Pro	48	144
Ovis	MedXpertQA	18	33
Ovis	All	164	406
Qwen	VQA-RAD	222	418
Qwen	MMMU-Pro	314	792
Qwen	MedXpertQA	202	410
Qwen	All	738	1356

B NAIVE BAYES TRAINING DETAILS

For the Naive Bayes text-confound baseline, we train a Laplace-smoothed binary multinomial model over lexical indicators extracted from model responses. Each example is represented by the set of unigrams and bigrams present in the response, using binary presence/absence features rather than counts. Class scores are computed as

$$\log p(y) + \sum_{g \in x} \log p(g \mid y),$$

with add-one smoothing applied to both class priors and feature likelihoods. As in probe training, examples with fewer than 10 response tokens are discarded.

For each model/setting/benchmark configuration, we run five seeds. In the all-examples setting, we first construct a class-balanced and benchmark-balanced subset, targeting 500 examples per class when available, and then use an 80/20 class-stratified train/test split. In the contrastive setting, we train on the full contrastive pool for the selected benchmark(s) and evaluate on a class-balanced, benchmark-stratified subset drawn from all-examples responses after removing any overlap with contrastive training questions.

We report mean accuracy across the five seeds. We additionally report cross-benchmark transfer by training on one benchmark and testing on another under the same balancing protocol (see below).

C NAIVE BAYES CROSS-BENCHMARK TRANSFER

Table 6: Naive Bayes classifier transfer metrics.

Model (Setting)	vqa→mmmu	vqa→medx	mmmu→vqa	mmmu→medx	medx→vqa	medx→mmmu
Ovis (Contrastive)	52.00%	53.33%	73.62%	16.67%	50.43%	53.83%
Ovis (All-Examples)	54.66%	49.47%	53.10%	50.00%	52.48%	52.05%
Qwen (Contrastive)	51.26%	46.87%	41.84%	49.48%	54.29%	51.36%
Qwen (All-Examples)	46.34%	47.45%	48.66%	51.31%	49.71%	56.92%

D PHRASE-GROUP CONFOUNDS

Table 7: Phrase-group class separators for Ovis. Cells show mirage-like example match rate minus non-mirage-like example match rate.

(a) Ovis Contrastive

Phrase Group	VQA-RAD	MMMU-Pro	MedXpertQA
reasoning_scaffold	-8.2%	+0.0%	-11.1%
image_grounding	-2.0%	+4.2%	+11.1%
uncertainty	-8.2%	-8.3%	-11.1%
hedging	-2.0%	-16.7%	+0.0%
answer_boilerplate	-12.2%	+4.2%	+0.0%
radiology_terms	+0.0%	+4.2%	-11.1%

(b) Ovis All-Examples

Phrase Group	VQA-RAD	MMMU-Pro	MedXpertQA
reasoning_scaffold	-15.1%	-32.8%	-11.0%
image_grounding	+8.7%	+7.1%	+10.6%
uncertainty	-22.6%	-3.7%	-9.8%
hedging	-22.4%	-7.6%	-7.1%
answer_boilerplate	-6.4%	+11.7%	-0.1%
radiology_terms	-3.9%	+10.2%	-1.0%

Table 8: Phrase-group class separators for Qwen. Cells show mirage-like example match rate minus non-mirage-like example match rate.

(a) Qwen Contrastive

Phrase Group	VQA-RAD	MMMU-Pro	MedXpertQA
reasoning_scaffold	+0.0%	-1.9%	+5.9%
image_grounding	+1.8%	-0.6%	-4.0%
uncertainty	+0.9%	+0.0%	-1.0%
hedging	+3.6%	-4.5%	+1.0%
answer_boilerplate	+0.9%	+0.6%	-5.9%
radiology_terms	+0.0%	+2.5%	+2.0%

(b) Qwen All-Examples

Phrase Group	VQA-RAD	MMMU-Pro	MedXpertQA
reasoning_scaffold	-4.0%	+3.6%	+7.8%
image_grounding	+3.0%	-3.8%	-32.0%
uncertainty	-10.1%	-0.1%	+0.8%
hedging	-14.3%	+8.4%	+4.9%
answer_boilerplate	+0.8%	+14.7%	-8.6%
radiology_terms	+0.0%	+2.6%	+0.8%

E MLP PROBE RESULTS

Table 9: Probe strategy performance on MLP-target activations.

Strategy	VQA-RAD	MMMU-Pro	MedXpertQA	All
Ovis – Contrastive				
LogReg	**76.33%**	61.67%	63.33%	72.00%
MLP	**71.33%**	60.00%	66.67%	69.33%
Concat	68.33%	49.00%	53.33%	**70.67%**
Diff	88.33%	**91.33%**	58.33%	89.00%
Ovis – All Examples				
LogReg	**71.63%**	67.82%	61.90%	69.96%
MLP	71.63%	65.52%	**85.71%**	70.37%
Concat	63.83%	**66.67%**	28.57%	60.49%
Diff	**95.04%**	86.21%	71.43%	90.95%
Qwen – Contrastive				
LogReg	**74.00%**	59.67%	60.00%	59.67%
MLP	**70.33%**	60.33%	59.67%	62.33%
Concat	**62.67%**	56.67%	47.33%	57.00%
Diff	**93.33%**	78.00%	80.00%	81.67%
Qwen – All Examples				
LogReg	59.13%	63.08%	**64.23%**	56.99%
MLP	71.36%	**75.65%**	68.60%	72.91%
Concat	67.14%	**71.88%**	64.25%	69.41%
Diff	**95.77%**	80.00%	76.81%	84.93%

F COMPUTE RESOURCES

All experiments are run using one or more NVIDIA RTX A6000s, each with 50 gigabytes of VRAM. Full Ovis runs on residual or additional targets (MLP, attention heads, post-attention) take around 12 hours to run with tasks orchestrated across two GPUs. Qwen runs require around 24 hours of activation extraction across four GPUs followed by 12 hours of probing orchestrated across two. All experiments (except for the Qwen activation extraction) could plausibly be run with much less powerful GPUs.

G Expanded Contrastive Logistic Regression Experiment

To reinforce our main finding that contrastive logistic regression probing reveals that mirage mechanisms are linearly represented in image-present latent space, we rerun our contrastive logistic regression probing experiment to include a total of 5 models and 5 datasets across residual streams, attention layers, individual attention heads, and MLP outputs. We use an experimental setup identical to that described in Section 3.5, except we only use the *text_nonspecial_mean* aggregation, which generally performs best, in order to keep the experiment tractable. Across all models and all benchmarks, we find that our claim of non-trivial linear decodability holds true, confirming the trend presented in Table 1. Results for residual and attention targets are presented in the tables below.

Table 10: **Expanded residual probe results.**

Model	VQA-RAD	MMMU-Pro	MedXpertQA	ChartQA	A-OKVQA	All
Ovis2.5-2B	72.20%	65.00%	66.00%	63.40%	57.80%	60.00%
Qwen3-VL-32B-Instruct	74.60%	60.00%	60.40%	67.00%	64.00%	61.00%
InternVL3.5-4B-HF	66.60%	64.60%	73.75%	60.60%	60.20%	59.60%
MiMo-VL-7B-RL	66.20%	58.20%	54.40%	63.80%	62.80%	58.80%
Gemma-3-4B-IT	68.80%	63.40%	77.78%	63.60%	58.00%	66.00%

Table 11: **Expanded attention probe results.**

Model	VQA-RAD	MMMU-Pro	MedXpertQA	ChartQA	A-OKVQA	All
Ovis2.5-2B	77.33%	70.33%	71.67%	72.33%	59.33%	64.33%
Qwen3-VL-32B-Instruct	76.67%	66.00%	69.67%	72.00%	67.00%	64.33%
InternVL3.5-4B-HF	70.33%	67.33%	85.42%	70.67%	66.33%	64.00%
MiMo-VL-7B-RL	74.00%	64.00%	60.33%	68.67%	65.67%	64.67%
Gemma-3-4B-IT	73.00%	66.67%	83.33%	62.67%	69.67%	68.67%

H Causal Main Table Experiments

In addition to the correlational results presented in Table 1, we conduct three additional experiments to demonstrate that mirage mechanisms are causally active in image-present latent space.

Steering with Identified Probe Directions. We inject the previously identified Ovis VQA-RAD image-present contrastive logistic probe direction at the layer 25 residual stream during both image-present and image-absent generation, sweeping both signs at plus/minus 4 training-projection SDs. Specifically, we evaluate 45 held-out base questions, 24 with images and 21 without, spanning mirage-like, non-mirage-like, and neutral labels. We find that steering produces a large, sign-reversing change in missing-image/refusal-token log probability (toward-mirage: -0.22 nats, away-from-mirage: +0.23 nats), indicating that the direction is causally implicated in the model's ability to provide confident, image-grounded responses.

Steering with Image-Present DIM Probes. For both Ovis and Qwen, we fit a difference-in-means (DIM) probe to the direction between confident no-image mirages and explicit no-image refusals, selecting steering layers and coefficients on a validation dataset. More specifically, we use all of the available yes/no VQA-RAD questions labeled mirage-like or non-mirage-like, employ a 60/20/20 train/valid/test split, fit probes on the train data, select layers and coefficients based on the number of answer flips or refusals steering with a particular strength/layer configuration achieves relative to a random matched-norm direction on the validation data, and then evaluate the best configuration on the test data. We also fit a DIM probe between labeled with-image examples at the chosen layer (using the same strength for steering). Using the image-absent probe, we steer image-absent mirage-like test examples toward refusal and steer image-absent non-mirage-like test examples toward mirage. We additionally try steering the image-absent mirage-like generations toward refusal using the image-present probe. Results are presented in Table 12, with percentages indicating behavioral flip rates. Randomly selected and orthogonal control directions perform significantly worse and fail to replicate the DIM probes' effects (see the parenthetical values).

Table 12: **DIM probe steering behavioral flip rates.**

Model	Mirage→Refuse	Refuse→Mirage	Mirage→Refuse (Image Probe)
Ovis	41% (0%)	7% (0%)	24%
Qwen	14% (0%)	6% (2%)	15%

Again, the results support the notion that causal aspects of mirage mechanisms are linearly encoded in model latent space, with steering along image-absent probe directions both inventing and mitigating mirages. Moreover, the demonstration that image-present probes causally transfer to the image-absent setting, mitigating mirages at similar rates to image-absent probes despite distribution shift, further supports the idea that image-present probe-identified directions are causally implicated in VLM mirage.

Identification of a Mirage Circuit. We use a two-stage autonomous circuit discovery pipeline to identify a late-layer 29 attention head circuit responsible for enabling mirage behavior in Ovis. First, using a dataset of 32 image-absent refusal/mirage pairs from all three benchmarks, EAP-IG (Hanna et al., 2024) ranks residual-stream edges by integrating how moving each edge from its mirage-side activation toward its refusal-side donor changes pre-annotated visual-reasoning versus evidence-unavailability token likelihoods. We use generic refusal, answer content, and final-answer tokens as nuisance controls. Subsequent ACDC (Conmy et al., 2023) pruning and an exact-head screen supply an initial 66-head circuit. Second, using a separate dataset of 24 questions where the image-absent and image-present reasoning and final responses are strongly contradictory, a second round of ACDC prunes the circuit to 29 heads by preserving bidirectional transfer between text-only-prior and image-added-claim token cohorts while bounding broad KL divergence. We test this circuit on image-absent (60 contrastive pairs) and image-present mirage-like (9 contrastive pairs) datasets by patching the selected heads' activations and ablating the heads altogether. Results are presented in the tables below. No tested control is able to reproduce the presented effects. All entries are mean token log-prob changes in nats.

Table 13: **Image-absent patching results.**

Image-Absent Patch	Donor-Distinctive Tokens
Honest Patched into Mirage	+0.0811
Mirage Patched into Honest	+0.0686

Thus, inserting an honest state into a mirage receiver increases distinctive missing-image-evidence language, while reversing the patch increases false image-grounded reasoning tokens.

Table 14: **Image-present mirage-like patching results.**

Patch Into Image-Present Mirage	No-Image Prior	Image-Added Evidence	Missing Evidence Language
Honest No-Image Donor	+0.1491	−0.5213	+0.2927
Mirage No-Image Donor	+0.2160	−0.5214	−0.0075
Zero All Heads	−0.4361	+0.2168	+0.7948

Both no-image donors disengage the receiver from image evidence while increasing the probability of no-image prior tokens. Crucially, patching in an honest donor selectively largely increases missing-evidence token probability. Zero ablation substantially reduces no-image prior token mass while significantly increasing missing-evidence token mass.

These results indicate that the VLM mirage mechanism is facilitated by a visual-grounding reasoning circuit that can become corrupted and populated by no-image priors, thus enabling false, ungrounded reasoning that is presented as grounded visual analysis. Patching in no-image states, regardless of whether they are mirage-like or not, reduces weight on visual priors because this circuit does truly process some visual information. Patching in the no-image states increases no-image prior weight because this corrupted circuit can also carry no-image priors. This is further supported by the no-image patching experiment, where the same circuit clearly processes false image-grounded

reasoning tokens. However, when patching in a non-mirage example, we see that missing-evidence token weight increases substantially, indicating that this circuit is sometimes able to properly calculate that no true visual information is available. Perhaps most convincingly, when the circuit is entirely ablated, no-image prior probability mass massively drops while missing-visual-evidence probability mass significantly increases, indicating that when active, this circuit facilitates both the false visual-grounding and the false no-image prior reasoning that define mirage behavior.

I Annotation Scheme Validation

We design our automated annotation scheme through iterative refinement against human judgment and cosine threshold sweeps. To validate our final labels, we annotate 100 response pairs, each consisting of responses generated with and without an image. Of these, 50 pairs come from Ovis and 50 from Qwen, with 20 pairs sampled from each of the five benchmarks and labels balanced across the sample. We also measure the percent change in total labels as we vary the cosine threshold.

Across the 100 pairs, marking all ambiguous cases as incorrect, we find a human-system agreement rate of 78%, which falls within the common acceptable range for such heuristic-based annotation systems used in the literature (Dammu et al., 2024; Han et al., 2024; Kaur et al., 2026). This agreement rate suggests that our labeling procedure does not introduce a significant amount of noise into our pipeline, further increasing confidence in our main results.

Percent change in labels across all Ovis and Qwen generations as we vary the cosine similarity threshold are presented in the table below. We find that both decreasing the threshold or increasing it slightly causes little meaningful change, supporting the robustness of our annotation pipeline.

Table 15: Cosine threshold label sensitivity. Cells show the percentage of labels that change relative to the default threshold.

Cosine Threshold	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90
Labels Changed	4.20%	3.84%	3.25%	2.10%	0.00%	3.29%	8.33%	14.53%	19.92%

J Causal Two Mechanism Experiment

Following a similar procedure to that described in the first section of Appendix H, we employ Ovis logistic regression probe directions for steering at plus/minus 4 training-projection SDs. In particular, we use the MedXpertQA and combined directions to steer 50 and 57 held-out MedXpertQA responses, respectively, and annotated the steered outputs using the same Claude-Sonnet-5 procedure described in Section 4.3 to check for behavioral flips. We observed no clean steering-induced behavioral flips. We then try steering 50 held-out MedXpertQA examples with the combined probe direction while tracking log probability weight assigned to visual-reference vocabulary before and after steering. We find no particularly strong sign-reversing change, and a random control direction reproduces the minimal effect we do observe. Accordingly, as steering along this direction appears to have no particularly strong causal effect whatsoever, instead inducing small amounts of random noise, we believe that it is a post-hoc readout direction for the two mechanisms, rather than a component of any larger causal process.